PATENT
Attorney Docket No. 3495.0010-01

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:

Marc Alizon, et al.

Serial No.: 07/158,652 — Group Art Unit: 1805

Filed: February 22, 1988 — Examiner: J. Railey

For: CLONED DNA SEQUENCES RELATED TO THE GENOMIC RNA OF LYMPHADENOPATHY ASSOCIATED VIRUS (LAV) AND PROTEINS ENCODED BY SAID LAV GENOMIC RNA

Honorable Commissioner of Patents
and Trademarks
Washington, D.C. 20231

Sir:

CLAIM FOR PRIORITY

Under the provisions of Section 119 of 35 U.S.C., applicants hereby claim the benefit of the filing date of Great Britain Application No. 84 29099, filed November 16, 1984, for the above identified United States Patent Application.

In support of applicants' claim for priority, filed herewith is one certified copy of GB 84 29099.

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N.W.
WASHINGTON, DC 20005
1-202-408-4000

If there are any fees due in connection with the filing of this Paper, please charge such fees to our Deposit Account No. 06-0916.

Respectfully submitted,

FINNEGAN, HENDERSON, FARABOW, GARRETT & DUNNER

By: _____
Michele M. Schafer
Reg. No. 34,717

Date: October 21, 1993

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N.W.
WASHINGTON, DC 20005
1-202-408-4000

THE PATENT OFFICE,
25 SOUTHAMPTON BUILDINGS,
LONDON.

I, the undersigned, being an officer duly authorised in accordance with Section 62(3) of the Patents and Designs Act 1907, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents, has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or the inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Please turn over

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.

Witness my hand this
21st day of OCTOBER 19 85

COC.1

16 NOV 1984

PATENTS ACT 1977

PATENTS FORM No. 1/77 (Revised 1982)
(Rules 16, 19)

19/11/84 B3662 PAT*** 10.0

The Comptroller
The Patent Office
25 Southampton Buildings
London, WC2A 1AY

1984
8429099

Fee: £10.00

REQUEST FOR GRANT OF A PATENT

8429099

THE GRANT OF A PATENT IS REQUESTED BY THE UNDERSIGNED ON THE BASIS OF THE PRESENT APPLICATION

I Agent's Reference JJD/EAF/26804

II Title of Invention CLONED DNA SEQUENCES RELATED TO THE GENOMIC RNA OF LYMPHADENOPATHY-ASSOCIATED VIRUS (LAV) AND PROTEINS ENCODED BY SAID LAV GENOMIC RNA.

III Applicant or Applicants *(See note 2)*

Name (First or only applicant) INSTITUT PASTEUR

Country FRANCE State ADP Code No.

Address 25-28 Rue du Dr. Roux,
75724 Paris Cedex 15, France.

Name (of second applicant, if more than one) Centre National de la Recherche Scientifique

Country FRANCE State

Address 15 Quay Anatole France
75007 Paris, France.

IV Inventor (see note 3)

or

(b) A statement on Patents Form No. 7/77 is/will be furnished

V Name of Agent (if any) *(See note 4)* Reddie & Grose ADP CODE NO

VI Address for Service (See note 5) 16 Theobalds Road
London WC1X 8PL

VII Declaration of Priority (See note 6)

Country	Filing date	File number

THE PATENT OFFICE
16 NOV 1984
25 SOUTHAMPTON BUILDINGS
LONDON WC2A 1AY

VIII The Application claims an earlier date under Section 8(3), 12(6), 15(4), 37(4) (See note 7)

Section No.

Earlier application or patent number and filing date

IX Check List *(To be filled in by applicant or agent)*

A The application contains the following number of sheet(s)		B The application as filed is accompanied by:–	
1 Request	1 Sheet(s)	1 Priority document	No
2 Description	17 Sheet(s)	2 Translation of priority document	No
3 Claim(s)	2 Sheet(s)	3 Request for Search	No
4 Drawing(s)	26 Sheet(s)	4 Statement of Inventorship and Right to Apply	No
5 Abstract	0 Sheet(s)	5	

X It is suggested that Figure No 1 of the drawings (if any) should accompany the abstract when published.

XI Signature *(See note 8)*

Reddie & Grose, Agents for the Applicant(s)

NOTES:

1. This form, when completed, should be brought or sent to the Patent Office together with the prescribed fee and two copies of the description of the invention, and of any drawings.
2. Enter the name and address of each applicant. Names of individuals should be indicated in full and the surname or family name should be underlined. The names of all partners in a firm must be given in full. Bodies corporate should be designated by their corporate name and the country of incorporation and, where appropriate, the state of incorporation within that country should be entered where provided. Full corporate details, eg "a corporation organised and existing under the laws of the State of Delaware, United States of America," trading styles, eg "trading as xyz company", nationality, and former names, eg "formerly [known as] ABC Ltd." are not required and should not be given. Also enter applicant(s) ADP Code No. (if known).
3. Where the applicant or applicants is/are the sole inventor or the joint inventors, the declaration (a) to that effect at IV should be completed, and the alternative statement (b) deleted. If, however, this is not the case the declaration (a) should be struck out and a statement will then be required to be filed upon Patent Form No 7/77.
4. If the applicant has appointed an agent to act on his behalf, the agent's name and the address of his place of business should be indicated in the spaces available at V and VI. Also insert agent's ADP Code No. (if known) in the box provided.
5. An address for service in the United Kingdom to which all documents may be sent must be stated at VI. It is recommended that a telephone number be provided if an agent is not appointed.
6. The declaration of priority at VII should state the date of the previous filing and the country in which it was made and indicate the file number, if available.
7. When an application is made by virtue of section 8(3), 12(6), 15(4), or 37(4) the appropriate section should be identified at VIII and the number of the earlier application or any patent granted thereon identified.
8. Attention is directed to rules 90 and 106 of the Patent Rules 1982.
9. Attention of applicants is drawn to the desirability of avoiding publication of inventions relating to any article, material or device intended or adapted for use in war (Official Secrets Acts, 1911 and 1920). In addition after an application for a patent has been filed at the Patent Office the comptroller will consider whether publication or communication of the invention should be prohibited or restricted under section 22 of the Act and will inform the applicant if such prohibition is necessary.
10. Applicants resident in the United Kingdom are also reminded that, under the provisions of section 23 applications may not be filed abroad without written permission or unless an application has been filed not less than six weeks previously in the United Kingdom for a patent for the same invention and no direction prohibiting publication or communication has been given or any such direction has been received.

Cloned DNA sequences related to the genomic RNA of lymphadenopathy-associated-virus (LAV) and proteins encoded by said LAV genomic RNA

The invention relates to cloned DNA sequences indistinguishable from genomic RNA and DNA of lymphadenopathy-associated virus (LAV), a process for their preparation and their uses. It relates more particularly to stable probes including a DNA sequence which can be used for the detection of the LAV virus or related viruses or DNA proviruses in any medium, particularly biological samples containing any of them. The invention also relates to polypeptides, whether glycosylated or not, encoded by said DNA sequences.

Lymphadenopathy-associated virus (LAV) is a human retrovirus first isolated from the lymph node of a homosexual patient with lymphadenopathy syndrome, frequently a prodrome or a benign form of acquired immune deficiency syndrome (AIDS). Subsequently other LAV isolates have been recovered from patients with AIDS or pre-AIDS. All available data are consistent with the virus being the causative agent of AIDS.

A method for cloning such DNA sequences has already been disclosed in British Patent Application Nr. 84 23858 filed on September 19, 1984. Reference is hereafter made to that application as concerns subject matter in common with the further improvements to the invention disclosed herein.

The present invention aims at providing additional new means which should not only also be useful for the detection of LAV or related viruses (hereafter more generally referred to as "LAV viruses"), but also have more versatility, particularly in detecting specific parts of the genomic DNA of said viruses whose expression products are not always directly detectable by immunological methods.

The present invention further aims at providing

polypeptides containing sequences in common with polypeptides encoded by the LAV genomic RNA. It relates even more particularly to polypeptides comprising antigenic determinants included in the proteins encoded and expressed by the LAV genome occuring in nature. An additional object of the invention is to further provide means for the detection of proteins related to LAV virus, particularly for the diagnosis of AIDS or pre-AIDS or, to the contrary, for the detection of antibodies against the LAV virus or proteins related therewith, particularly in patients afflicted with AIDS or pre-AIDS or more generally in asymtomatic carriers and in blood-related products. Finally the invention also aims at providing immunogenic polypeptides, and more particularly protective polypeptides for use in the preparation of vaccine compositions against AIDS or related syndroms.

The present invention relates to additional DNA fragments, hybridizable with the genomic RNA of LAV as they will be disclosed hereafter, as well as with additional cDNA variants corresponding to the whole genomes of LAV viruses. It further relates to DNA recombinants containing said DNAs or cDNA fragments.

The invention relates more particularly to a cDNA variant corresponding to the whole of LAV retroviral genomes, which is characterized by a series of restriction sites in the order hereafter (from the 5' end to the 3' end).

The coordinates of the successive sites of the whole LAV genome (restriction map) are indicated hereafter too, with respect to the Hind III site (selected as of coordinate 1) which is located in the R region. The coordinates are estimated with an accuracy of ± 200 bp :

Hind III	0
Sac I	50
Hind III	520
Pst I	800
Hind III	1 100

Bgl II	1 500
Kpn I	3 500
Kpn I	3 900
Eco RI	4 100
Eco RI	5 300
Sal I	5 500
Kpn I	6 100
Bgl II	6 500
Bgl II	7 600
Hind III	7 850
Bam HI	8 150
Xho I	8 600
Kpn I	8 700
Bgl II	8 750
Bgl II	9 150
Sac I	9 200
Hind III	9 250

Another DNA variant according to this invention optionally contains an additional Hind III approximately at the 5 550 coordinate.

Reference is further made to fig. 1 which shows a more detailed restriction map of said whole-DNA (λJ19).

An even more detailed nucleotidie sequence of a preferred DNA according to the invention is shown in fig. 4-12 hereafter.

The invention further relates to other preferred DNA fragments which will be referred to hereafter.

Additional features of the invention will appear in the course of the non-limitative disclosure of additional features of preferred DNAs of the invention, as well as of preferred polypeptides according to the invention. Reference will further be had to the drawings in which :

- fig. 1 is the restriction map of a complete LAV genome (clone λJ19) :
- figs. 2 and 3 show diagrammatically parts of the three

possible reading phases of LAV genomic RNA, including the open reading frames (ORF) apparent in each of said reading phases ;

- figs. 4-12 show the successive nucleotidic sequences of a complete LAV genome. The possible peptidic sequences in relation to the three possible reading phases related to the nucleotidic sequences shown are also indicated ;
- figs. 13-18 reiterate the sequence of part of the LAV genome containing the genes coding for the enveloppe proteins, with particular boxed peptidic sequences which corresponds to groups which normally carry glycosyl groups.

The sequencing and determination of sites of particular interest was carried out on a phage recombinant corresponding to AJ19 disclosed in the abovesaid British Patent application Nr. 84 23859. A method for preparing it is disclosed in that application.

The whole recombinant phage DNA of clone AJ19 (disclosed in the earlier application) was sonicated according to the protocol of DEININGER (1983), Analytical Biochem. 129, 216. the DNA was repaired by a Klenow reaction for 12 hours at 16°C. The DNA was electrophoresed through 0.8 % agarose gel and DNA in the size range of 300-800 bp was cut out and electroeluted and precipitated. Resuspended DNA (in 10 mM Tris, pH 8 ; 0,1 mM EDTA) was ligated into M13mp8 RF DNA (cut by the restriction enzyme SmaI and subsequently alkaline phosphated), using T4 DNA- and RNA-ligases (Maniatis T et al (1982) - Molecular cloning - Cold Spring Harbor Laboratory). An E. coli strain designated as TG1 was used for further study. This strain has the following genotype :

Δlac pro, supE, thi.F'tra036, proAB, $lacI^q$, ZΔM15,r^-

This E. coli TG1 strain has the peculiarity of enabling recombinants to be recognized easily. The blue colour of the cells transfected with plasmids which did

not recombine with a fragment of LAV DNA is not modified. To the contrary cells transfected by a recombinant plasmid containing a LAV DNA fragment yield white colonies. The technique which was used is disclosed in Gene (1983), 26, 101.

This strain was transformed with the ligation mix using the Hanahan method (Hanahan D (1983) J. Mol. Biol. 166, 557). Cells were plated out on tryptone-agarose plate with IPTG and X-gal in soft agarose. White plaques were either picked and screened or screened directly in situ using nitrocellulose filters. Their DNAs were hybridized with nick-translated DNA inserts of pUC18 Hind III subclones of λJ19. this permitted the isolation of the plasmids or subclones of λ which are identified in the table hereafter. In relation to this table it should also be noted that the designation of each plasmid is followed by the deposition number of a cell culture of E. coli TGI containing the corresponding plasmid at the "Collection Nationale des Cultures de Micro-organismes" (C.N.C.M.) of the Pasteur Institute in Paris, France. A non-transformed TGI cell line was also deposited at the C.N.C.M. under Nr. I-364. All these deposits took place on November 15, 1984. The sizes of the corresponding inserts derived from the LAV genome have also been indicated.

TABLE

Essential features of the recombinant plasmids

- pJ19 - 1 plasmid (I-365) 0.5 kb

Hind III - Sac I - Hind III

- pJ19 - 17 plasmid (I-367) 0.8 kb

Hind III - Pst 1 - Hind III

- pJ19 - 6 plasmid (I-366) 1.5 kb

Hind III (5')
Bam HI
Xho I
Kpn I
Bgl II
Sac I (3')
Hind III

- pJ19-13 plasmid (I-368) 6.7 kb

Hind III (5')
Bgl II
Kpn I
Kpn I
Eco RI
Eco RI
Sal I
Kpn I
Bgl II
Bgl II
Hind III (3')

M : méthionine
W : tryptophan
F : phenylalanine
Y : tyrosine
L : leucine
V : valine
I : isoleucine
G : glycine
T : thréonine
S : serine
E : glutamic acid
D : Aspartic acid
N : asparagine
Q : glutamine
P : proline.

The asterik signs "*" correspond to stop codons (i.e. TAA, TAG and TGA).

Starting above the first line of the DNA nucleotidic sequence of fig. 4 the three reading phases are respectively marked "1", "2", "3", on the left handside of the drawing. The same relative presentation of the three theoritical reading phases is then used all over the successives lines of the LAV nucleotidic sequence.

Figs. 2 and 3 provide a diagrammatized representation of the lengths of the successive open reading frames corresponding to the successive reading phases (also referred to by numbers "1", "2" and "3" appearing in the left handside part of fig. 2). The relative positions of these open reading frames (ORF) with respect to the nucleotidic structure of the LAV genome is referred to by the scale of numbers representative of the respective positions of the corresponding nucleotides in the DNA sequence. The vertical bars correspond to the positions of the corresponding stop codons.

1) <u>The "gag gene" (or ORF-gag)</u>

The "gag gene" codes for core proteins.

Particularly it appears that a genomic fragment (ORF-gag) thought to code for the core antigens including the p25, p18 and p13 proteins is located between nucleotidic position 236 (starting with 5' CTA GCG GAG 3') and nucleotidic position 1759 (ending by CTCG TCA CAA 3'). The structure of the peptides or proteins encoded by parts of said ORF is deemed to be that corresponding to phase 2.

The methionine aminoacid "M" coded by the ATG at position 280-282 is the probable initiation methionine of the gag protein precursor. The end of ORF-gag and accordingly of gag protein appears to be located at position 1759.

The beginning of p25 protein, thought to start by a P-I-V-Q-N-I-Q-G-Q-M-V-H .... aminoacid sequence is thought to be coded for by the nucleotidic sequence CCTATA..., starting at position 658.

Hydrophilic peptides in the gag open reading frame are identified hereafter. They are defined starting from aminoacid 1 = Met (M) coded by the ATG starting from 280-2 in the LAV DNA sequence.

Those hydrophilic peptides are

12-32	aminoacids	inclusive
37-46	"	"
49-79	"	"
88-153	"	"
158-185	"	"
178-188	"	"
200-220	"	"
226-234	"	"
239-284	"	"
288-331	"	"
352-381	"	"
377-390	"	"
399-432	"	"
437-484	"	"
492-498	"	"

The invention also relates to any combination of these peptides.

2) <u>The "pol gene" (or ORF-pol)</u>

Figs. 4-12 also show that the DNA fragments extending from nucleotidic position 1555 (starting with 5'TTT TTT ....3' to nucleotidic position 5088 is thought to correspond to the pol gene. The polypeptidic structure of the corresponding polypeptides is deemed to be that corresponding to phase 1. It stops at position 4563 (end by 5'G GAT GAG GAT 3').

These genes are thought to code for the virus polymerase or reverse transcriptase.

3) <u>The envelope gene (or ORF-env)</u>

The DNA sequence thought to code for envelope proteins is thought to extend from nucleotidic position 5670 (starting with 5'AAA GAG GAG A....3') up to nucleotidic position 8132 (ending by ....A ACT AAA GAA 3'). Polypeptidic structures of sequences of the envelope protein correspond to those read according to the "phase 3" reading phase.

The start of env transcription is thought to be at the level of th ATG codon at positions 5691-5693.

Additional feature of the envelope protein coded by the env genes appear on figs. 13-18. These are to be considered as paired figs. 13 and 14 ; 15 and 16 ; 17 and 18 respectively.

It is to be mentioned that because of format difficulties.

Fig. 14 overlaps to some extent with fig. 13.

Fig. 16 overlaps to some extent with fig. 15.

Fig. 18 overlaps to some extent with fig. 17.

Thus for instance figs. 13 and 14 must be considered together. Particularly the sequence shown on the first line on the top of fig. 13 overlaps with the sequence shown on the first line on the top of fig. 14. In other words the starting of the reading of the successive

sequences of the env gene as represented in figs. 13-18 involves first reading the first line at the top of fig. 13 then proceeding further with the first line of fig. 14. One then returns to the beginning of the second line of fig. 13, then again further proceed with the reading of the second line of page 14, etc... The same observations then apply to the reading of the paired figs. 15 and 16, and paired figs. 17 and 18, respectively.

The locations of neutralizing epitopes are further apparent in figs. 13-18, reference is more particularly made to the boxed groups of three letters included in the aminoacid sequences of the envelope proteins (reading phase 3) which can be designated generally by the formula N-X-S or N-X-T, wherein X is any other possible aminoacid. Thus the initial protein product of the env gene in a glycoprotein of molecular weight in excess of 91,000. These groups are deemed to generally carry glycosylated groups. These N-X-S and N-X-T groups with attached glycosylated groups form together hydrophylic regions of the protein and are deemed to be located at the periphery of and to be exposed outwardly with respect to the normal conformation of the proteins. Consequently they are considered as being epitopes which can efficiently be brought into play in vaccine compositions.

The invention thus concerns with more particularity peptide sequences included in the env-proteins and excizable therefrom (or having the same aminoacid structure), having sizes not exceeding 200 aminoacids.

Preferred peptides of this invention (referred to hereafter as a, b, c, d, e, f) are deemed to correspond to those encoded by the nucleotide sequences which extend respectively between the following positions :

a) from about 6095 to about 6200
b) " " 6260 " " 6310
c) " " 6390 " " 6440
d) " " 6465 " " 6620

e) " " 6860 " " 6930

f) " " 7535 " " 7630

Other hydrophilic peptides in the env open reading frame are identified hereafter. they are defined starting from

aminoacid 1 = lysine (K) coded by the AAA at position 5670-2 in the LAV DNA sequence.

These hydrophilic peptides are

8-23 aminoacids inclusive

63-78 " "

82-90 " "

97-123 " "

127-183 " "

197-201 " "

239-294 " "

300-327 " "

334-381 " "

397-424 " "

466-500 " "

510-523 " "

551-577 " "

594-603 " "

621-630 " "

657-679 " "

719-758 " "

780-803 " "

The invention also relates to any combination of these peptides.

4) <u>The other ORF</u>

The invention further concerns DNA sequences which provide open reading frames defined as ORF-Q, ORF-R and as "1", "2", "3", "4", "5", the relative position of which appears more particularly in figs. 2 and 3.

These ORFs have the following locations :

ORF-Q phase 1 start 4478 stop 5086

ORF-R " 2 " 8249 " 8896

ORF-1	"	1	"	5029	" 5316
ORF-2	"	2	"	5273	" 5515
ORF-3	"	1	"	5383	" 5616
ORF-4	"	2	"	5519	" 5773
ORF-5	"	1	"	7986	" 8279

The LTR (long terminal repeats) can be defined as lying between position 8560 and position 180 (end extending over position 9097/1). As a matter of fact the end of the genome is at 9097 and, because of the LTR structure of the retrovirus, links up with the beginning of the sequence :

Hind III

CTCAATAAAGCTTGCCTTG

9097 1

The invention concerns more particularly all the DNA fragments which have been more specifically referred to hereabove and which correspond to open reading frames. It will be understood that the man skilled in the art will be able to obtain them all, for instance by cleaving an entire DNA corresponding to the complete genome of a LAV species, such as by cleavage by a partial or complete digestion thereof with a suitable restriction enzyme and by the subsequent recovery of the relevant fragments. The different DNAs disclosed in the earlier mentioned British Application can be resorted to also as a source of suitable fragments. The techniques disclosed hereabove for the isolation of the fragments which were then included in the plasmids referred to hereabove and which were then used for the DNA sequencing can be used.

Of course other methods can be used. Some of them have been examplified in the earlier British Application. reference is for instance made to the following methods.

a) DNA can be transfected into mammalian cells with appropriate selection markers by a variety of techniques, calcium phosphate precipitation, polyethylene

glycol, protoplast-fusion, etc..

b) DNA fragments corresponding to genes can be cloned into expression vectors for E. coli , yeast- or mammalian cells and the resultant proteins purified.

c) The provival DNA can be "shot-gunned" (fragmented) into procaryotic expression vectors to generate fusion polypeptides. Recombinant producing antigenically competent fusion proteins can be identified by simply screening the recombinants with antibodies against LAV antigens .

The invention also relates more specifically to cloned probes which can be made starting from any DNA fragment according to this invention, thus to recombinant DNAs containing such fragments, particularly any plasmids amplifiable in procaryotic or eucaryotic cells and carrying said fragments.

Using the cloned DNA fragments as a molecular hybridization probe - either by marking with radionucleotides or with fluorescent reagents - LAV virion RNA may be detected directly in the blood, body fluids and blood products (e.g. of the antihemophylic factors such as Factor VIII concentrates) and vaccines, i.e. hepatitis B vaccine. It has already been shown that whole virus can be detected in culture supernatants of LAV producing cells. A suitable method for achieving that detection comprises immobilizing virus onto said a support e.g. nitrocellulose filters, etc., disrupting the virion and hybridizing with labelled (radiolabelled or "cold" fluorescent- or enzyme-labelled) probes. Such an approach has already been developed for Hepatitis B virus in peripheral blood (according to SCOTTO J. et al. Hepatology (1983), 3, 379-384).

Probes according to the invention can also be used for rapid screening of genomic DNA derived from the tissue of patients with LAV related symptoms, to see if the proviral DNA or RNA is present in host tissue and other

tissues.

A method which can be used for such screening comprise the following steps : extraction of DNA from tissue, restriction enzyme cleavage of said DNA, electrophoresis of the fragments and Southern blotting of genomic DNA from tissues, subsequent hybridization with labelled cloned LAV provival DNA. Hybridization in situ can also be used.

Lymphatic fluids and tissues and other non-lymphatic tissues of humans, primates and other mammalian species can also be screened to see if other evolutionnary related retrovirus exist. The methods referred to hereabove can be used, although hybridization and washings would be done under non stringent conditions.

The DNA according to the invention can be used also for achieving the expression of LAV viral antigens for diagnostic purposes.

The invention also relates to the polypeptides themselves which can be expressed by the different DNAs of the inventions, particularly by the ORFs or fragments thereof, in appropriate hosts, particularly procaryotic or eucaryotic hosts, after transformation thereof with a suitable vector previously modified by the corresponding DNAs.

These polypeptides can be used as diagnostic tools, particularly for the detection of antibodies in biological media, particularly in sera or tissues of persons afflicted with pre-AIDS or AIDS, or simply carrying antibodies in the absence of any apparent disorders. Conversely the different peptides according to this invention can be used themselves for the production of antibodies, preferably monoclonal antibodies specific of the different peptides respectively. For the production of hybridomas secreting said monoclonal antibodies conventional production and screening methods are used. These monoclonal antibodies, which themselves are part of

the invention then provide very useful tools for the identification and even determination of relative proportions of the different polypeptides or proteins in biological samples, particularly human samples containing LAV or related viruses.

Thus all of the above peptides can be used in diagnostics as sources of immunogens or antigens free of viral particles, produced using non-permissive systems, and thus of little or no biohazard risk.

The invention further relates to the hosts (procaryotic or eucaryotic cells) which are transformed by the above mentioned recombinants and which are capable of expressing said DNA fragments.

Finally it also relates to vaccine compositions whose active principle is to be constituted by any of the expressed antigens, i.e. whole antigens, fusion polypeptides or oligopeptides in association with a suitable pharmaceutical or physiologically acceptable carrier.

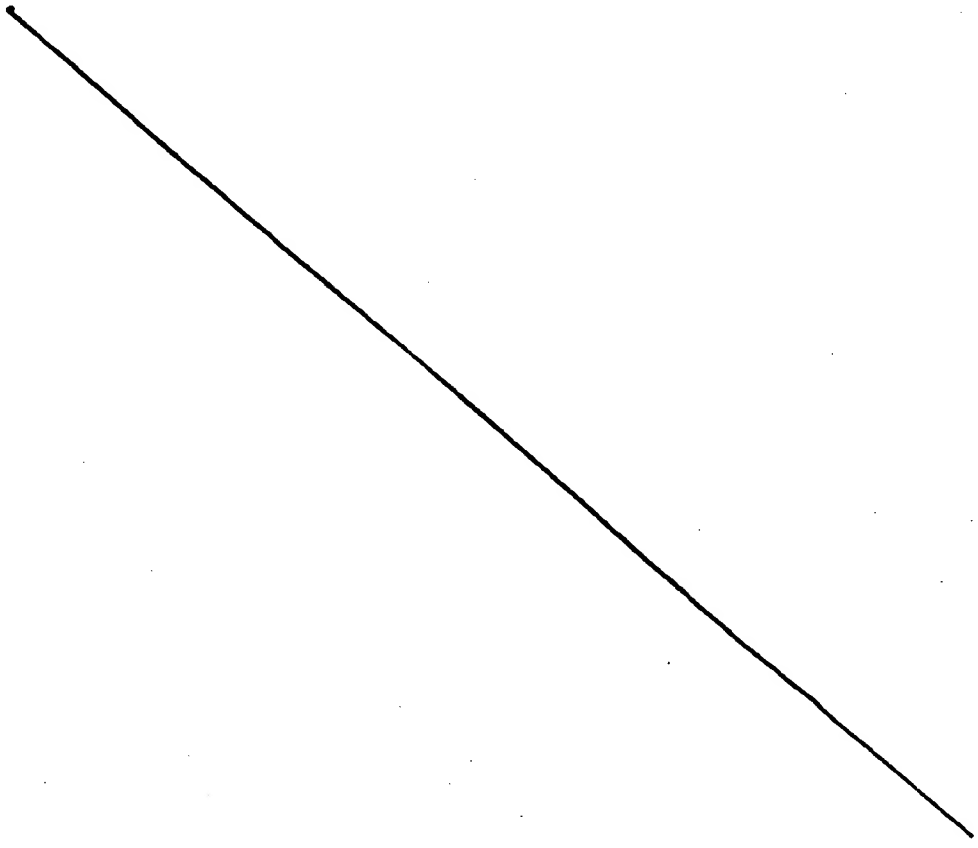
Preferably the active principles to be considered in that field consist of the peptides containing less than 250 aminoacid units, preferably less than 150 as deducible for the complete genomes of LAV, and even more preferably those peptides which contain one or more groups selected from N-X-S and N-X-T as defined above. Preferred peptides for use in the production of vaccinating principles are peptides (a) to (f) as defined above. By way of example having no limitative character, there may be mentioned that suitable dosages of the vaccine compositions are those which enable administration to the host, particularly human host ranging from 10 to 500 micrograms per kg, for instance 50 to 100 micrograms per kg.

For the purpose of clarity figs. 19 to 26 are added, reference may be made thereto in case of difficulties of reading blurred parts of figs. 4 to 12.

Needless to say that figs. 19-28 are merely a reiteration of the whole DNA sequence of the LAV genome.

Finally the invention also concerns vectors for the transformation of eucaryotic cells of human origin, particularly lymphocytes, the polymerases of which are capable of recognizing the LTRs of LAV. Particularly said vectors are characterized by the presence of a LAV LTR therein, said LTR being then active as a promoter enabling the efficient transcription and translation in a suitable host of the above defined, of a DNA insert coding for a determined protein placed under its controls.

Needless to say that the invention extends to all variants of genomes and corresponding DNA fragments (ORFs) having substantially equivalent properties, all of said genomes belonging to retroviruses which can be considered as equivalents of LAV.



CLAIMS

1. A DNA fragment of LAV extending from nucleotide position 236 to nucleotide position 1759.

2. A DNA fragment of LAV extending from nucleotide position 1555 to nucleotide position 5086.

3. A DNA fragment of LAV extending from nucleotide position 5870 to nucleotide position 8132.

4. A vector containing a DNA fragment according to any of claims 1 to 3.

5. Peptide corresponding to any of those encoded by the nucleotide sequences which extend respectively between the following positions :

a) from about 6095 to about 6200
b) " " 6260 " " 6310
c) " " 6390 " " 6440
d) " " 6485 " " 6620
e) " " 6860 " " 6930
f) " " 7535 " " 7630

6. Peptide characterized by a sequence of amino-acids deducible from LAV DNA the terminal aminoacids of which extend between the following positions with respect to the lysine (position 1) coded by the AAA at position 5870-5872 in the LAV DNA.

8-23 aminoacids inclusive
63-78 " "
82-90 " "
97-123 " "
127-183 " "
197-201 " "
239-294 " "
300-327 " "
334-381 " "
397-424 " "
466-500 " "
510-523 " "
551-577 " "

594-603 " "
621-630 " "
657-679 " "
719-758 " "
780-803 " "

or any combination of these peptides.

7. Peptide corresponding to the aminoacid sequences deducible from LAV DNA and the terminal aminoacids of which are positionned at the positions hereafter counted from the Met at position 1 coded by the ATG sequence at nucleotide positions 260-2 :

12-32 aminoacids inclusive
37-46 " "
49-79 " "
88-153 " "
158-165 " "
178-188 " "
200-220 " "
226-234 " "
239-264 " "
288-331 " "
352-361 " "
377-390 " "
398-432 " "
437-484 " "
492-498 " "

and combination of said peptides.

8. Diagnostic means containing any of the DNA fragments of any of claims 1 to 3.

9. Diagnostic means containing any of the peptides of any of claims 4 to 6.

10. Vaccine compositions containing any of the peptides according to any of claims 4 to 6 in association with a pharmaceutical vehicle.

End of transmission

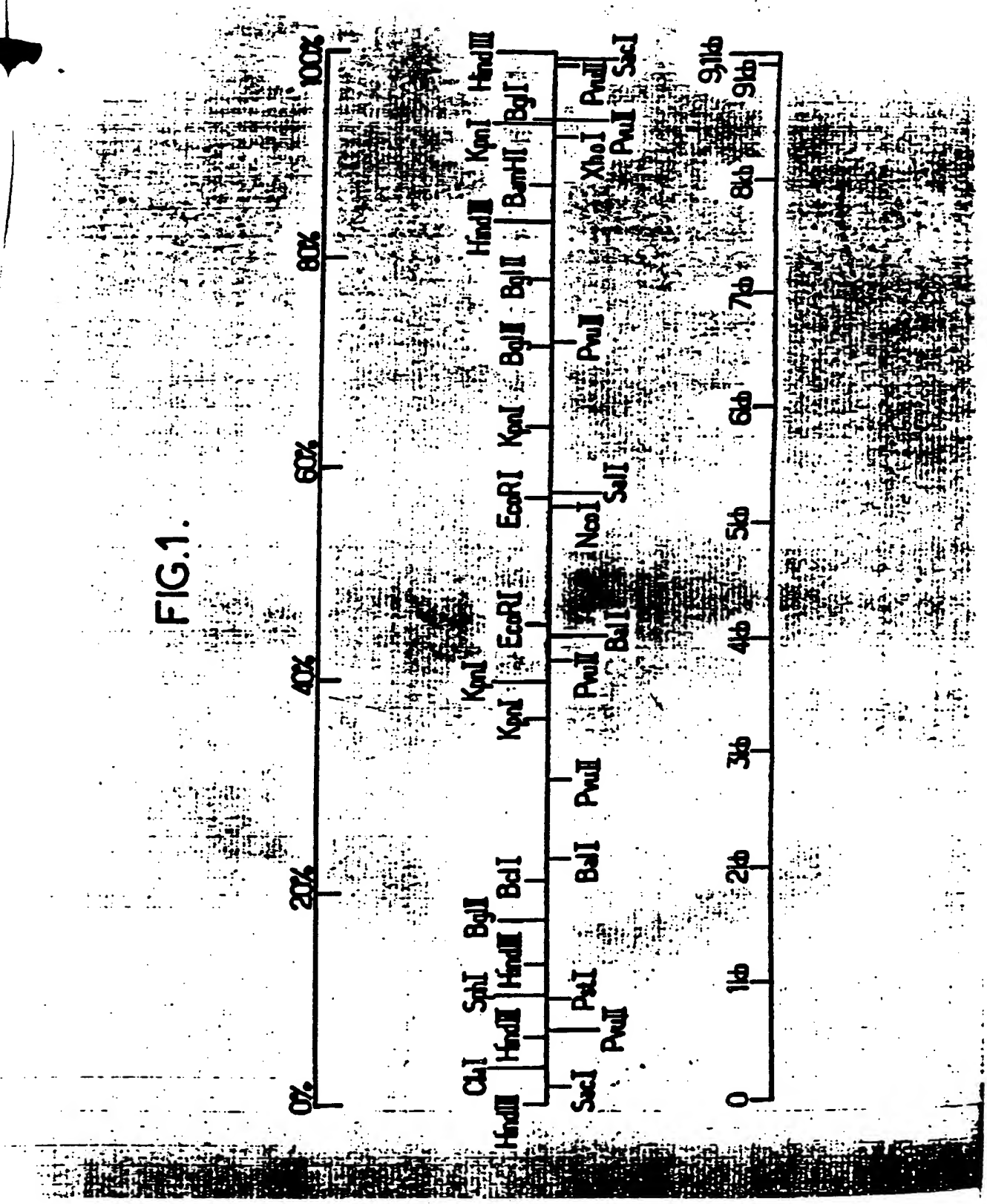


FIG.1.

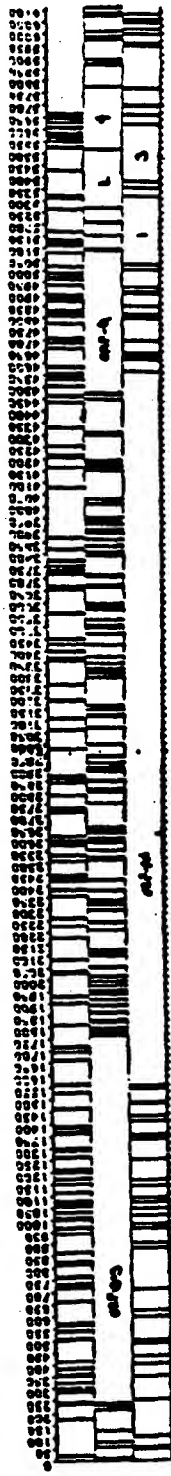


Fig. 2

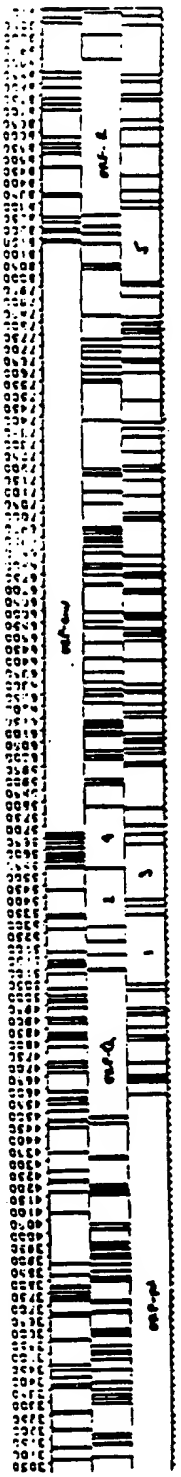


Fig. 3

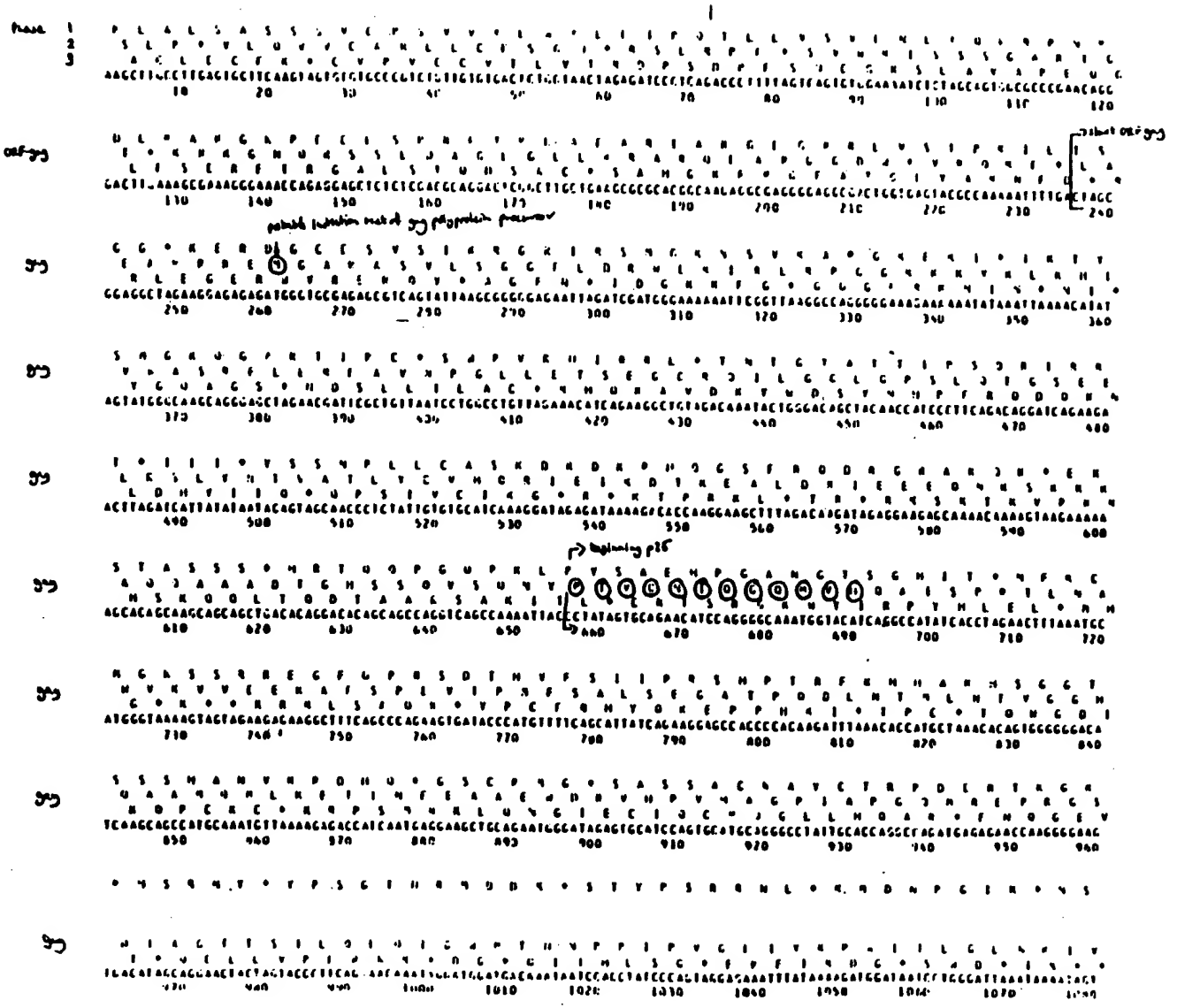


Fig 4

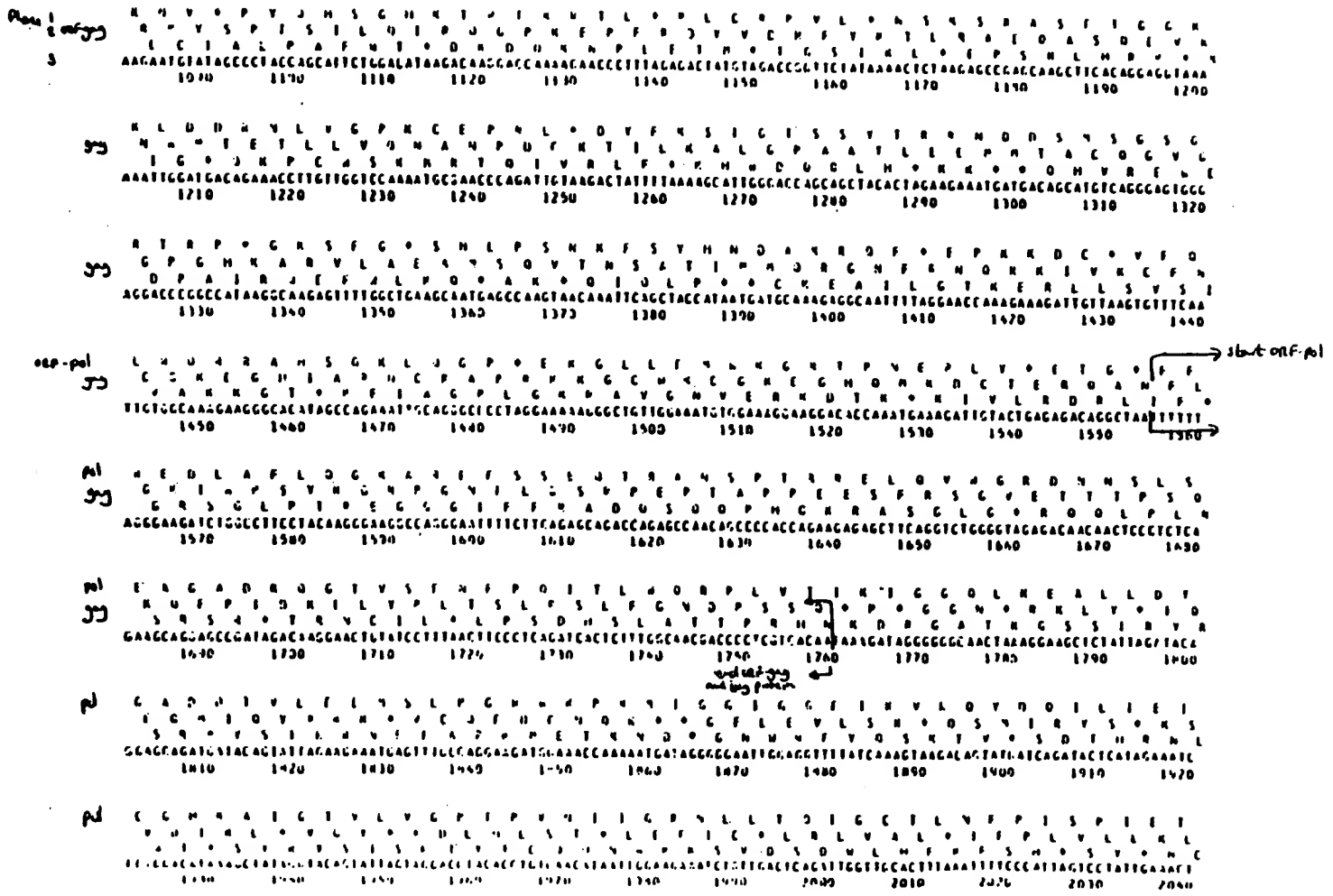


Fig. 5

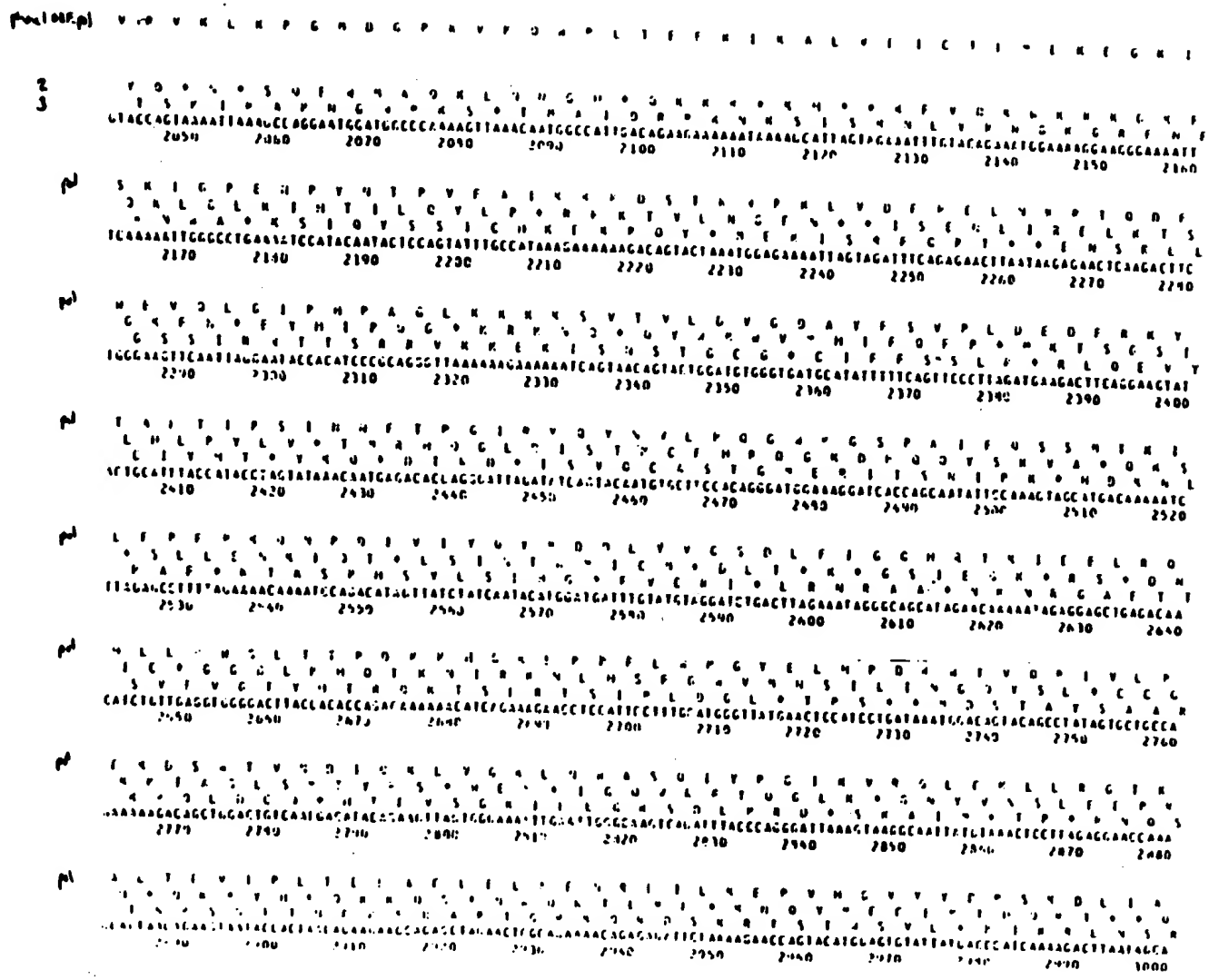
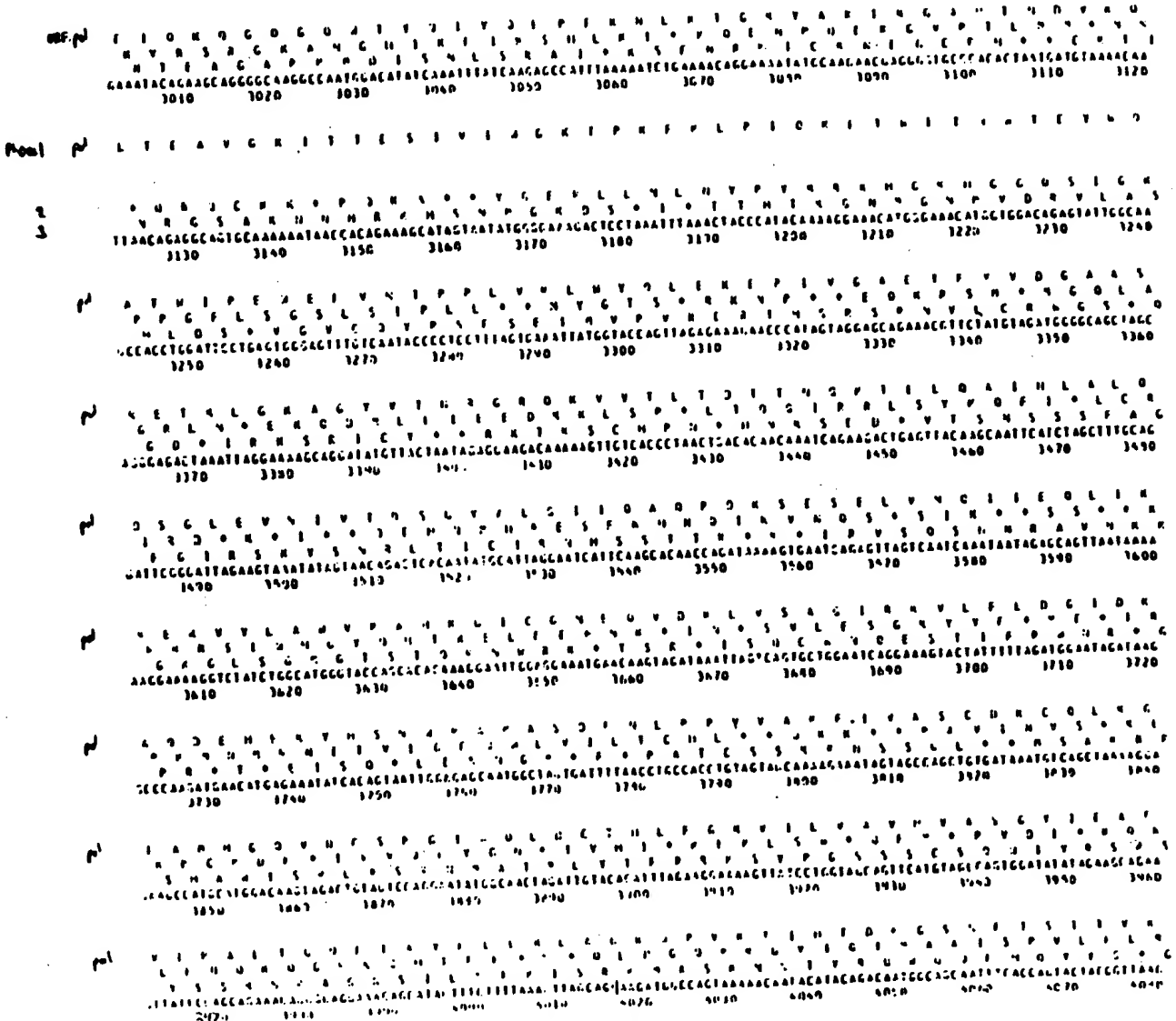


Fig 6



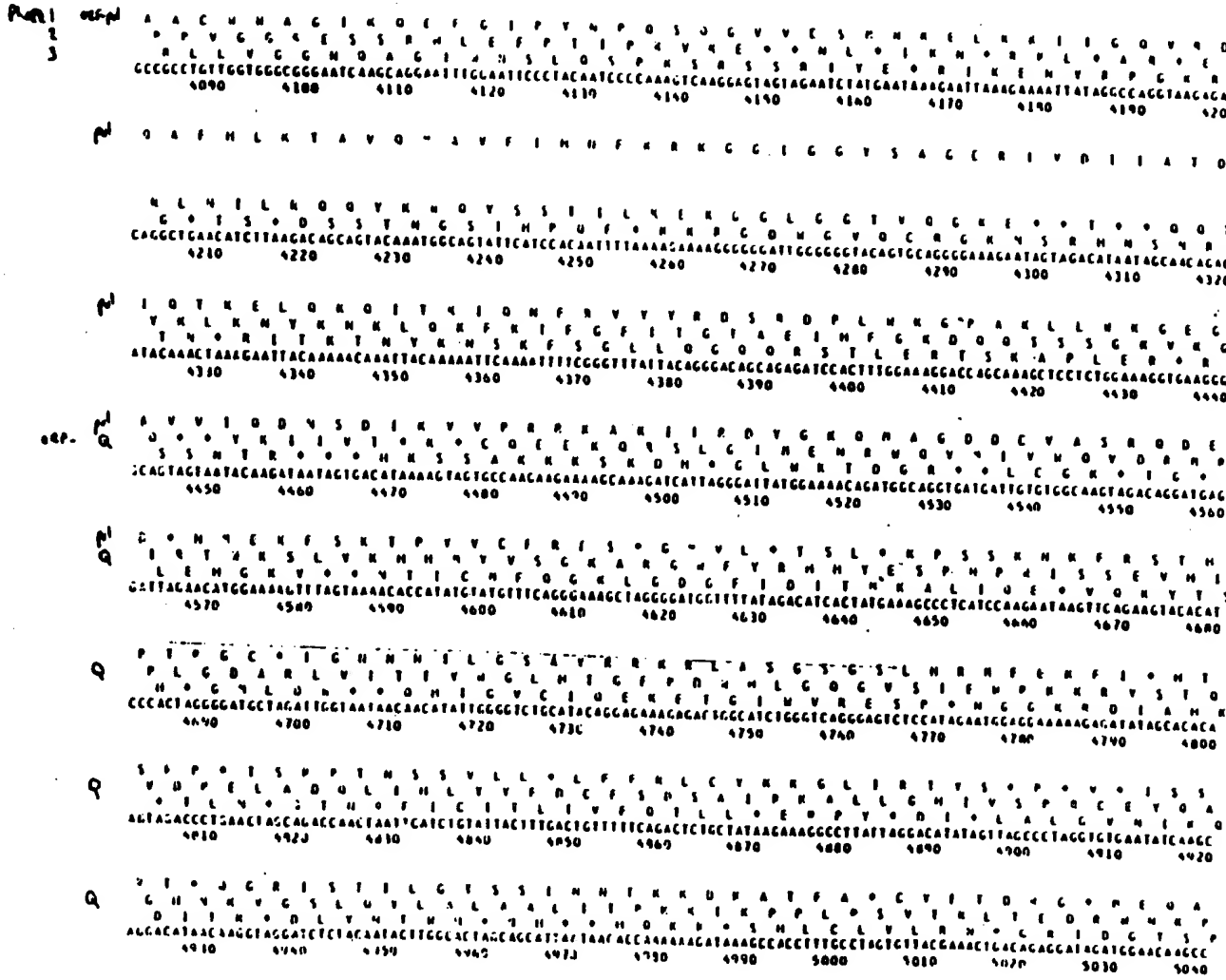


Fig. 8

16 NOV 84- 29099

DFA

9/26

16 NOV 94 - 29099

DFA

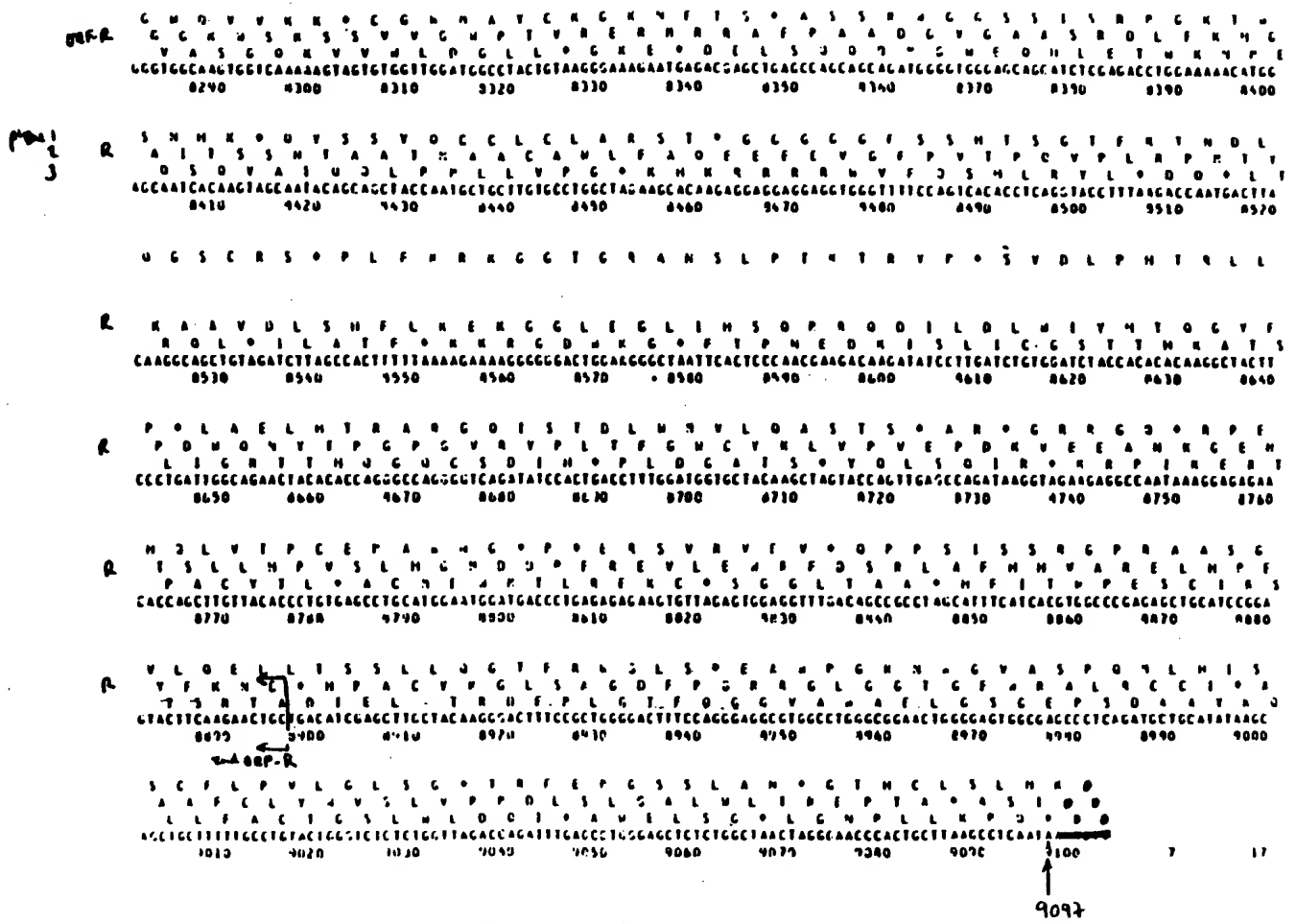
10/26

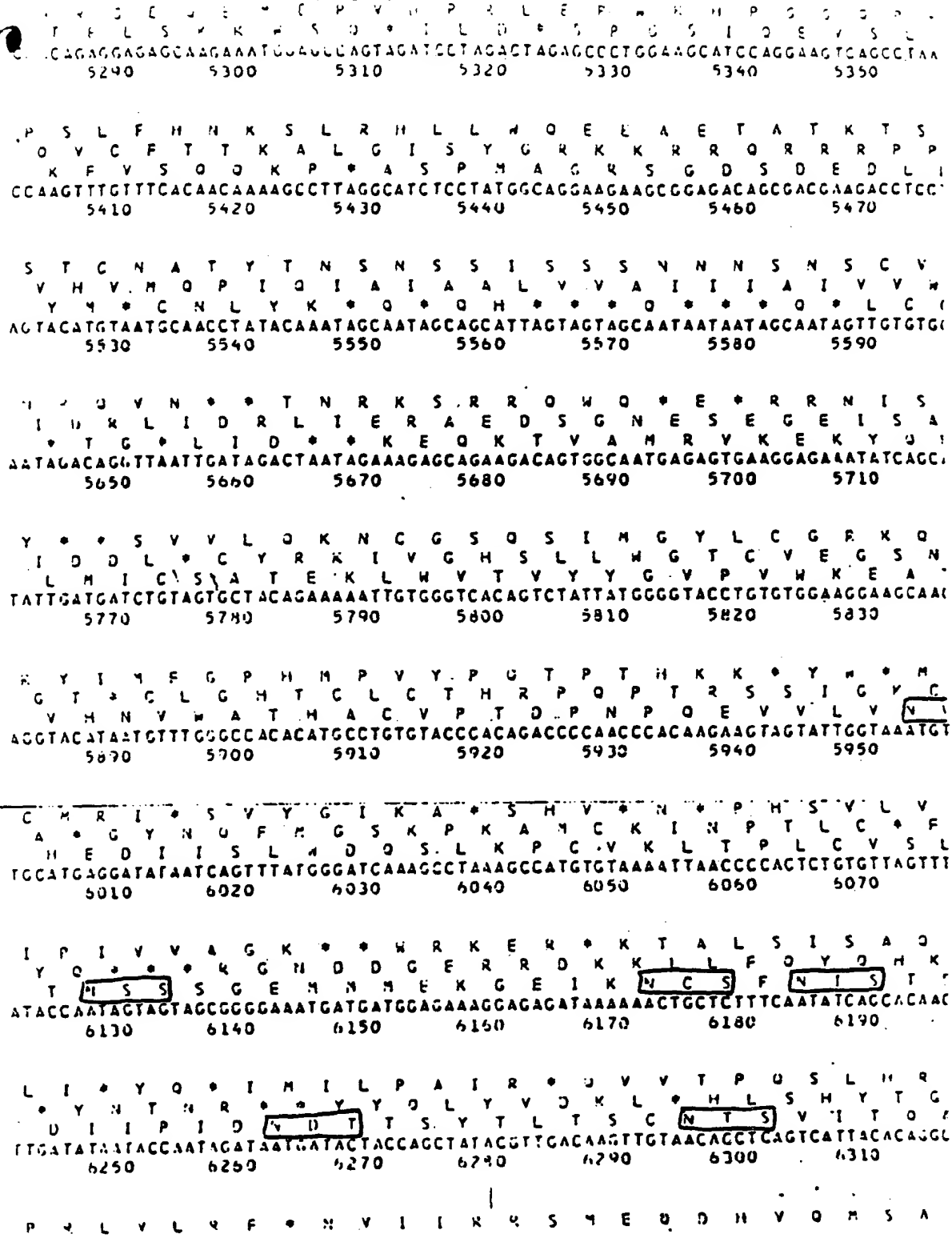
Fig 10

16 NOV. 84- 29099

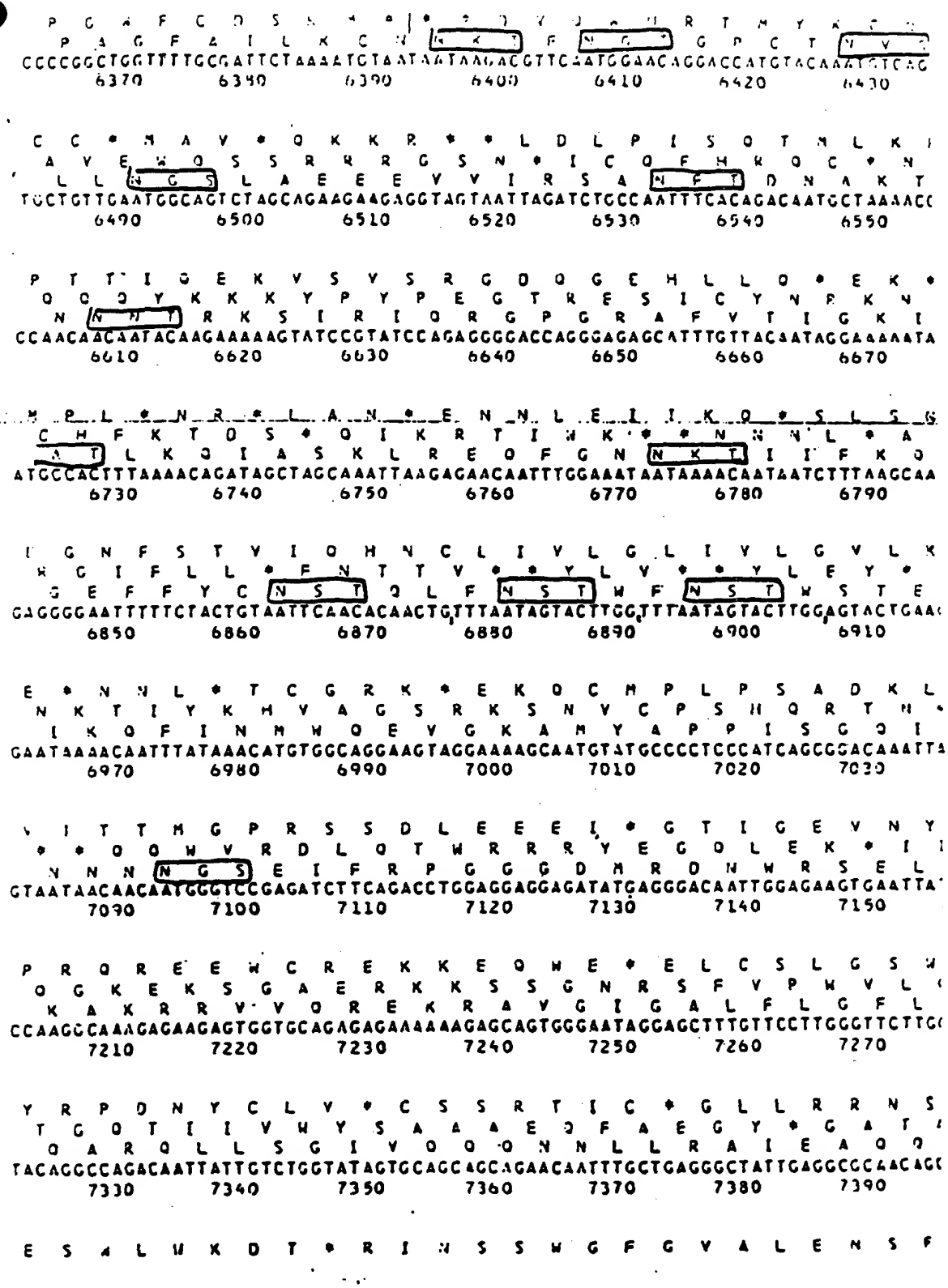
DFA 11/26

Fig 11





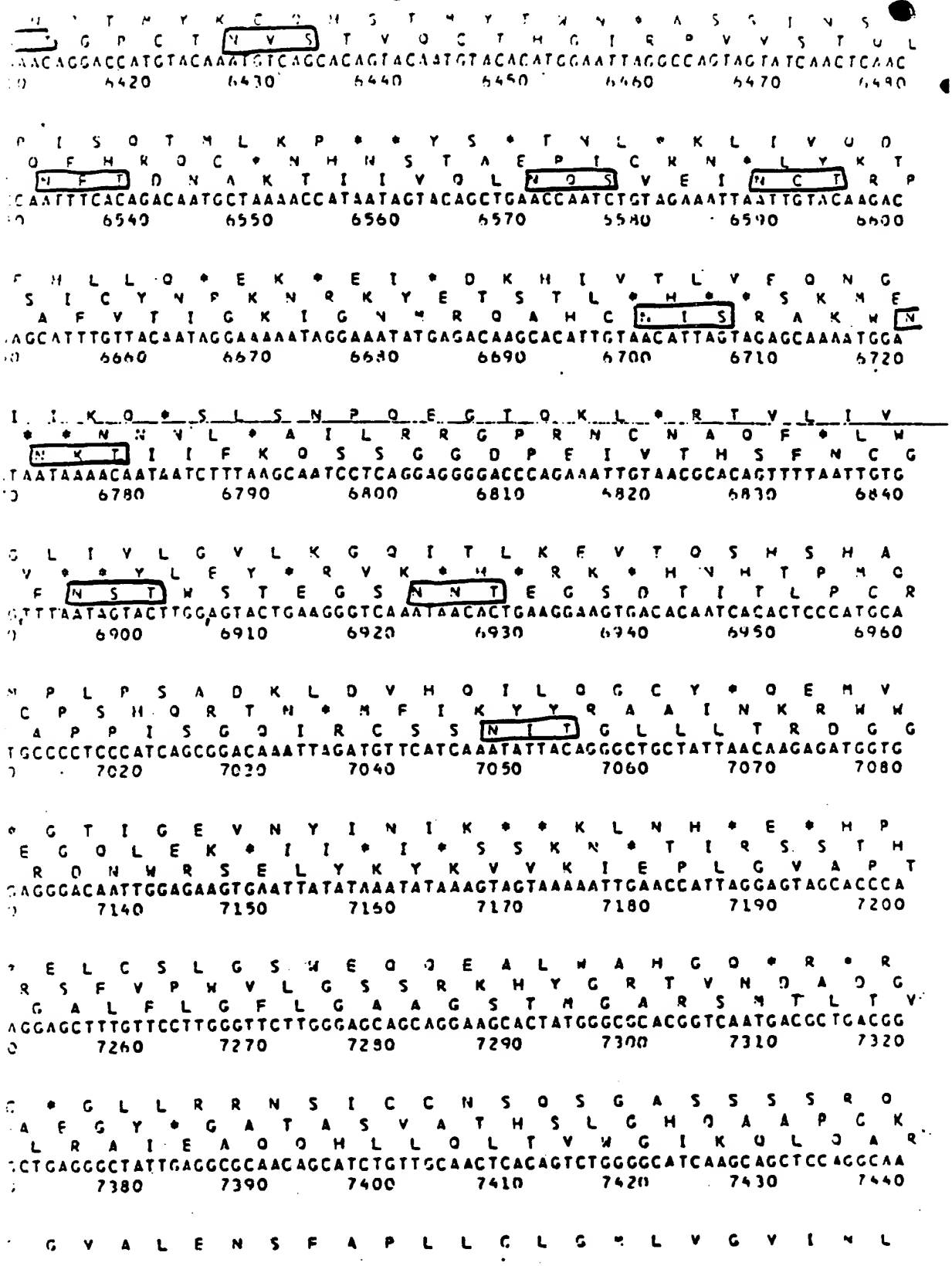
```
  .  .  S  Q  P  K  T  A  C  T  T  C  Y  C  K  K  C  C  F  H  C
  Q  .  V  S  L  K  L  L  V  P  L  A  I  V  K  S  V  A  F  I  A
CAGCAAGTCAGCCTAAAACTGCTTGTACCACTTGCTATTGTAAAAAGTGTTGCTTTCATTG
)       5350       5360       5370       5380       5390       5400

  A  T  K  T  S  S  P  Q  S  D  S  S  S  F  S  I  K  A  V  S
Q  R  R  R  P  P  Q  G  S  Q  T  H  Q  V  S  L  S  K  Q  *  V
  S  D  E  D  L  L  K  A  V  R  L  I  K  F  L  Y  Q  S  S  K  *
AGCGACGAAGACCTCCTCAAGGCAGTCAGACTCATCAAGTTTCTCTATCAAAGCAGTAAGT
Q       5470       5480       5490       5500       5510       5520

  S  N  S  C  V  V  H  S  N  H  R  I  *  E  N  I  K  T  K  K
I  A  I  V  V  W  S  I  V  I  I  E  Y  R  K  I  L  R  Q  R  K
  *  Q  *  L  C  G  P  *  *  S  *  N  I  G  K  Y  *  D  K  E  K
TAGCAATAGTTGTGTGGTCCATAGTAATCATAGAATATAGGAAAATATTAAGACAAAGAAA
)       5590       5600       5610       5620       5630       5640

  R  R  N  I  S  T  C  G  D  G  G  G  N  G  A  P  C  S  L  G
  G  E  I  S  A  L  V  E  M  G  V  E  M  G  H  H  A  P  W  D
K  E  K  Y  Q  H  L  W  R  W  G  W  K  W  G  T  M  L  L  G  I
AAGGAGAAATATCAGCACTTGTGGAGATGGGGGTGGAAATGGGGCACCATGCTCCTTGGGA
)       5710       5720       5730       5740       5750       5760

  C  G  F  K  Q  P  P  L  Y  F  V  H  Q  M  L  K  H  M  I  Q
  V  E  G  S  N  H  H  S  I  L  C  I  R  C  *  S  I  *  Y  R
V  W  K  E  A  T  T  T  L  F  C  A  S  D  A  K  A  Y  D  T  E
TGTGGAAGGAAGCAACCACCACTCTATTTTGTGCATCAGATGCTAAAGCATATGATACAG
        5830       5840       5850       5960       5870       5880

  *  Y  W  *  M  *  Q  K  I  L  T  C  G  K  M  T  W  *  N  R
  S  I  G  K  C  D  R  K  F  *  H  V  E  K  *  H  G  R  T  D
V  V  L  Y  N  V  T  E  N  F  N  M  W  K  N  D  M  V  E  Q  M
TAGTATTGGTAAATGTGACAGAAAATTTTAACATGTGGAAAAATGACATGGTAGAACAGA
        5950       5960       5970       5980       5990       6000

  H  S  V  L  V  *  S  A  L  I  W  G  M  L  L  I  P  I  V  V
  T  L  C  *  F  K  V  H  *  F  G  E  C  Y  *  Y  Q  *  *  *
Q  L  C  V  S  L  K  C  T  D  L  G  N  A  T  N  T  N  S  S  N
CACTCTGTGTTAGTTTAAAGTGCACTGATTTGGGGAAATGCTACTAATACCAATAGTAGTA
)       6070       6080       6090       6100       6110       6120

  S  I  S  A  Q  A  *  E  V  R  C  P  K  N  M  H  F  F  I  N
  Q  Y  Q  H  K  H  K  R  *  G  A  E  R  I  C  I  F  L  *  T
F  N  I  S  T  S  I  R  G  K  V  Q  K  E  Y  A  F  F  Y  K  L
TCAATATCAGCACAAGCATAAGAGGTAAGGTGCAGAAAGAATATGCATTTTTTTATAAAC
        6190       6200       6210       6220       6230       6240

  Q  S  L  H  R  P  V  Q  R  Y  P  L  S  Q  F  P  Y  I  I  V
  S  H  Y  T  G  L  S  K  G  I  L  *  A  N  S  H  T  L  L  C
S  V  I  T  Q  A  C  P  K  V  S  F  E  P  I  P  I  H  Y  C  A
CAGTCATTACACAGGCCTGTCCAAAGGTATCCTTTGAGCCAATTCCCATACATTATTGTG
        6310       6320       6330       6340       6350       6360

  V  Q  M  S  A  Q  Y  N  V  H  *  F  L  G  Q  *  Y  Q  L  N
```

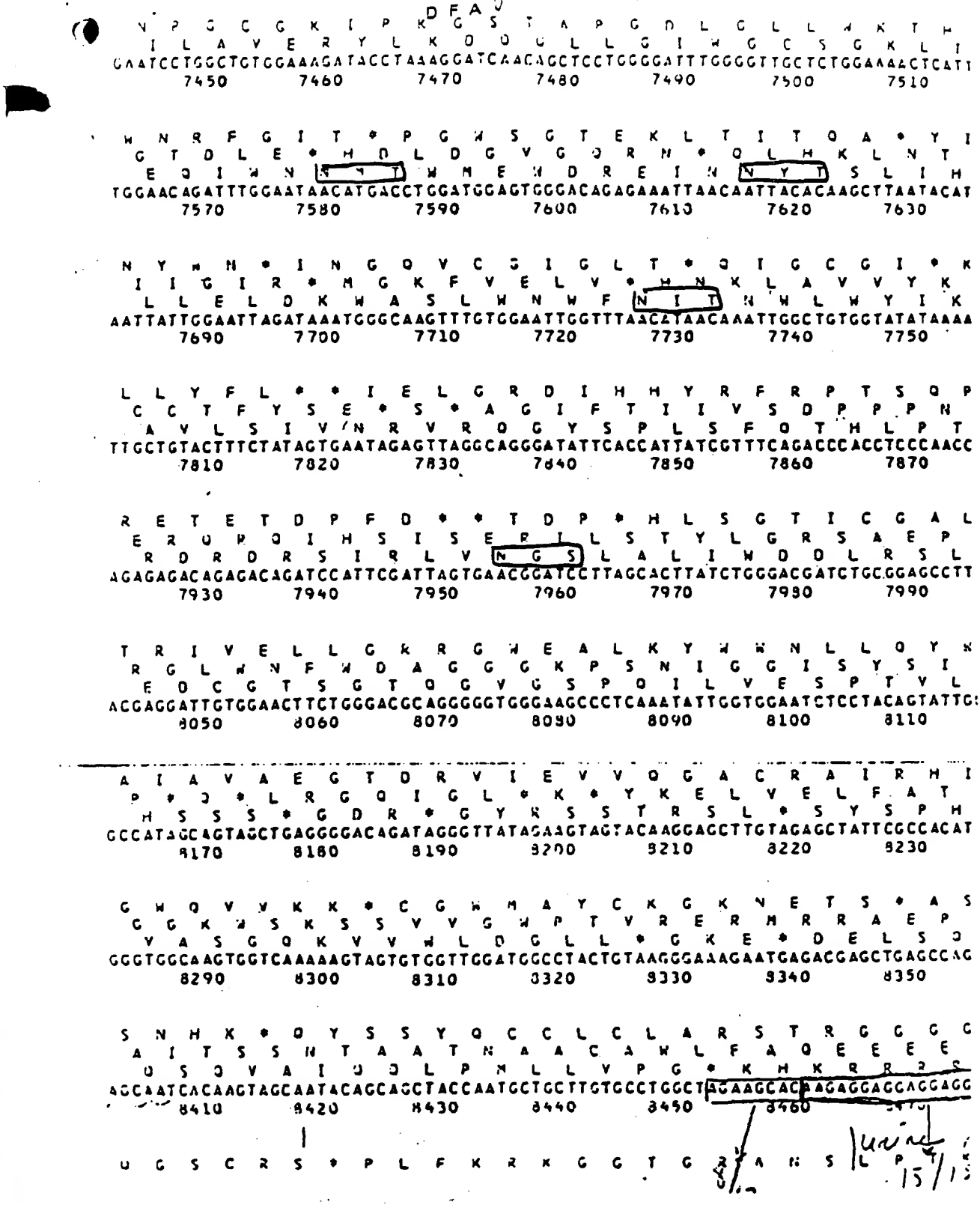

6370 6380 6390 6400 6410 6420 6430
6490 6500 6510 6520 6530 6540 6550
6610 6620 6630 6640 6650 6660 6670
6730 6740 6750 6760 6770 6780 6790
6850 6860 6870 6880 6890 6900 6910
6970 6980 6990 7000 7010 7020 7030
7090 7100 7110 7120 7130 7140 7150
7210 7220 7230 7240 7250 7260 7270
7330 7340 7350 7360 7370 7380 7390

16 NOV. 88 29299 16/26

Fig. 96

DFA




DFAS
7450 7460 7470 7480 7490 7500 7510
7570 7580 7590 7600 7610 7620 7630
7690 7700 7710 7720 7730 7740 7750
7810 7820 7830 7840 7850 7860 7870
7930 7940 7950 7960 7970 7980 7990
8050 8060 8070 8080 8090 8100 8110
8170 8180 8190 8200 8210 8220 8230
8290 8300 8310 8320 8330 8340 8350
8410 8420 8430 8440 8450 8460

16 NOV 84 - 28899

Fig 18

DFA

```
  N  K  T  H  L  H  H  C  C  A  L  E  C  *  L  E  *  *  I  S
  S  K  L  I  C  T  T  A  V  P  W  V  A  S  M  S  M  S  L
CTGGAAAACTCATTTGCACCACTGCTGTGCCTTGGAATGCTAGTTGGAGTAATAAATCTC
      7510       7520       7530       7540       7550       7560

  Q  A  *  Y  I  P  *  L  K  N  R  K  T  S  K  K  R  M  N  K
  K  L  N  T  F  L  N  *  R  I  A  K  P  A  R  K  E  *  T  R
  D  S  L  I  H  S  L  I  E  E  S  Q  N  Q  Q  E  K  N  E  Q  E
GAAGCTTAATACATTCCTTAATTGAAGAATCGCAAAACCAGCAAGAAAAGAATGAACAAG
      7630       7640       7650       7660       7670       7680

  C  G  I  *  K  Y  S  *  *  *  *  E  A  W  *  V  *  E  *  F
  V  V  Y  K  N  I  H  N  D  S  R  R  L  G  R  F  K  N  S  F
  W  Y  I  K  I  F  I  M  I  V  G  G  L  V  G  L/R/I  V  F
TGTGGTATATAAAAATATTCATAATGATAGTAGGAGGCTTGGTAGGTTTAAGAATAGTTT
      7750       7760       7770       7780       7790       7800

  P  T  S  Q  P  R  G  D  P  T  G  P  K  E  *  K  K  K  V  E
  P  P  P  N  P  E  G  T  R  Q  A  R  R  N  R  R  R  R  W  R
  T  H  L  P  T  P  R  G  P  D  R  P  E  G  I  E  E  E  G  G  E
CCCACCTCCCAACCCCGAGGGGACCCGACAGGCCCGAAGGAATAGAAGAAGAAGGTGGAG
      7870       7880       7990       7900       7910       7920

  I  C  G  A  L  C  L  F  S  Y  H  R  L  R  D  L  L  L  I  V
  S  A  E  P  C  A  S  S  A  T  T  A  *  E  T  Y  S  *  L  *
  L  R  S  L  V  P  L  Q  L  P  P  L  E  R  L  T  L  D  C  N
ATCTGCGGAGCCTTGTGCCTCTTCAGCTACCACCGCTTGAGAGACTTACTCTTGATTGTA
      7990       8000       8010       8020       8030       8040

  L  L  Q  Y  H  S  Q  E  L  K  N  S  A  V  S  L  L  N  A  T
  S  Y  S  I  G  V  R  N  *  R  I  V  L  L  A  C  S  M  P  Q
  P  T  Y  L  E  S  G  T  K  E  *  C  C  *  L  A  Q  C  H  S
TCCTACAGTATTGGAGTCAGGAACTAAAGAATAGTGCTGTTAGCTTGCTCAATGCCACA
      8110       8120       8130       8140       8150       8160

  A  I  R  H  I  P  R  R  I  R  Q  G  L  E  R  I  L  L  *  D
  L  F  A  T  Y  L  E  E  *  D  R  A  W  K  G  F  C  Y  K  M
  Y  S  P  H  T  *  K  N  K  T  G  L  G  K  D  F  A  I  R  W
CTATTCGCCACATACCTAGAAGAATAAGACAGGGCTTGGAAAGGATTTTGCTATAAGAT
      8230       8240       8250       8260       8270       8280

  T  S  *  A  S  S  R  A  G  G  S  S  I  S  R  P  G  K  T  W
  R  A  E  P  A  A  D  G  V  G  A  A  S  R  D  L  E  K  H  G
  E  L  S  Q  Q  Q  *  G  W  E  Q  M  L  E  T  W  K  N  M  E
CGAGCTGAGCCAGCAGCAGATGGGGTGGGAGCAGCATCTCGAGACCTGGAAAAACATGG
      8350       8360       8370       8390       8390       8400

  R  G  G  G  G  G  F  S  S  H  T  S  G  T  F  K  T  N  D  L
  E  E  E  E  V  G  F  P  V  T  P  Q  V  P  L  R  P  M  T  Y
  R  R  R  R  W  V  F  P  S  H  L  R  Y  L  *  D  Q  *  L  T
AGGAGGAGGAGGTGGGTTTTCCAGTCACACCTCAGGTACCTTTAAGACCAATGACTTA
      8470       8480       8490       8500       8510       8520

  S  L  P  T  K  T  R  Y  P  *  S  V  G  L  P  M  T  R  L  L
```

Unique Insert

15/15 B/L

DFA

Fig 19

```
        10         20         30         40         50         60
AAGCTTGCGT TGAGTGCTTC AAGTAGTGTG TGCCCGTCTG TTGTGTGACT CTGGTAACTA

        70         80         90        100        110        120
GAGATCCCTC AGACCCTTTT AGTCAGTGTG GAAAATCTCT AGCAGTGGCG CCCGAACAGG

       130        140        150        160        170        180
GACTTGAAAG CGAAAGGGAA ACCAGAGGAG CTCTCTCGAC GCAGGACTCG GCTTGCTGAA

       190        200        210        220        230        240
GCGCGCACGG CAAGAGGCGA GGGGAGGCGA CTGGTGAGTA CGCCAAAAAT TTTGACTAGC

       250        260        270        280        290        300
GGAGGCTAGA AGGAGAGAGA TGGGTGCGAG AGCGTCAGTA TTAAGCGGGG GAGAATTAGA

       310        320        330        340        350        360
TCGATGGGAA AAAATTCGGT TAAGGCCAGG GGGAAAGAAA AAATATAAAT TAAAACATAT

       370        380        390        400        410        420
AGTATGGGCA AGCAGGGAGC TAGAACGATT CGCTGTTAAT CCTGGCCTGT TAGAAACATC

       430        440        450        460        470        480
AGAAGGCTGT AGACAAATAC TGGGACAGCT ACAACCATCC CTTCAGACAG GATCAGAAGA

       490        500        510        520        530        540
ACTTAGATCA TTATATAATA CAGTAGCAAC CCTCTATTGT GTGCATCAAA GGATAGAGAT

       550        560        570        580        590        600
AAAAGACACC AAGGAAGCTT TAGACAAGAT AGAGGAAGAG CAAAACAAAA GTAAGAAAAA

       610        620        630        640        650        660
AGCACAGCAA GCAGCAGCTG ACACAGGACA CAGCAGCCAG GTCAGCCAAA ATTACCCTAT

       670        680        690        700        710        720
AGTGCAGAAC ATCCAGGGGC AAATGGTACA TCAGGCCATA TCACCTAGAA CTTTAAATGC

       730        740        750        760        770        780
ATGGGTAAAA GTAGTACAAG AGAAGGCTTT CAGCCCAGAA GTGATACCCA TGTTTTCAGC

       790        800        810        820        830        840
ATTATCAGAA GGAGCCACCC CACAAGATTT AAACACCATG CTAAACACAG TGGGGGGACA

       850        860        870        880        890        900
TCAAGCAGCC ATGCAAATGT TAAAAGAGAC CATCAATGAG GAAGCTGCAG AATGGGATAG

       910        920        930        940        950        960
AGTGCATCCA GTGCATGCAG GGCCTATTGC ACCAGGCCAG ATGAGAGAAC CAAGGGGAAG

       970        980        990       1000       1010       1020
TGACATAGCA GGAACTACTA GTACCCTTCA GGAACAAATA GGATGGATGA CAAATAATCC

      1030       1040       1050       1060       1070       1080
ACCTATCCCA GTAGGAGAAA TTTATAAAAG ATGGATAATC CTGGGATTAA ATAAAATAGT

      1090       1100       1110       1120       1130       1140
```

Fig 20

```
AAGAATGTAT AGCCCTACCA GCATTCTGGA CATAAGACAA GGACCAAAAG AACCCTTTAG

      1150        1160        1170        1180        1190        1200
AGACTATGTA GACCGGTTCT ATAAAACTCT AAGAGCCGAG CAAGCTTCAC AGGAGGTAAA

      1210        1220        1230        1240        1250        1260
AAATTGGATG ACAGAAACCT TGTTGGTCCA AAATGCGAAC CCAGATTGTA AGACTATTTT

      1270        1280        1290        1300        1310        1320
AAAAGCATTG GGACCAGCAG CTACACTAGA AGAAATGATG ACAGCATGTC AGGGAGTGGG

      1330        1340        1350        1360        1370        1380
AGGACCCGGC CATAAGGCAA GAGTTTTGGC TGAAGCAATG AGCCAAGTAA CAAATTCAGC

      1390        1400        1410        1420        1430        1440
TACCATAATG ATGCAAAGAG GCAATTTTAG GAACCAAAGA AAGATTGTTA AGTGTTTCAA

      1450        1460        1470        1480        1490        1500
TTGTGGCAAA GAAGGGCACA TAGCCAGAAA TTGCAGGGCC CCTAGGAAAA AGGGCTGTTG

      1510        1520        1530        1540        1550        1560
GAAATGTGGA AAGGAAGGAC ACCAAATGAA AGATTGTACT GAGAGACAGG CTAATTTTTT

      1570        1580        1590        1600        1610        1620
AGGGAAGATC TGGCCTTCCT ACAAGGGAAG GCCAGGGAAT TTTCTTCAGA GCAGACCAGA

      1630        1640        1650        1660        1670        1680
GCCAACAGCC CCACCAGAAG AGAGCTTCAG GTCTGGGGTA GAGACAACAA CTCCCTCTCA

      1690        1700        1710        1720        1730        1740
GAAGCAGGAG CCGATAGACA AGGAACTGTA TCCTTTAACT TCCCTCAGAT CACTCTTTGG

      1750        1760        1770        1780        1790        1800
CAACGACCCC TCGTCACAAT AAAGATAGGG GGGCAACTAA AGGAAGCTCT ATTAGATACA

      1810        1820        1830        1840        1850        1860
GGAGCAGATG ATACAGTATT AGAAGAAATG AGTTTGCCAG GAAGATGGAA ACCAAAAATG

      1870        1880        1890        1900        1910        1920
ATAGGGGGAA TTGGAGGTTT TATCAAAGTA AGACAGTATG ATCAGATACT CATAGAAATC

      1930        1940        1950        1960        1970        1980
TGTGGACATA AAGCTATAGG TACAGTATTA GTAGGACCTA CACCTGTCAA CATAATTGGA

      1990        2000        2010        2020        2030        2040
AGAAATCTGT TGACTCAGAT TGGTTGCACT TTAAATTTTC CCATTAGTCC TATTGAAACT

      2050        2060        2070        2080        2090        2100
GTACCAGTAA AATTAAAGCC AGGAATGGAT GGCCCAAAAG TTAAACAATG GCCATTGACA

      2110        2120        2130        2140        2150        2160
GAAGAAAAAA TAAAAGCATT AGTAGAAATT TGTACAGAAA TGGAAAAGGA AGGGAAAATT

      2170        2180        2190        2200        2210        2220
TCAAAAATTG GGCCTGAAAA TCCATACAAT ACTCCAGTAT TTGCCATAAA GAAAAAAGAC

      2230        2240        2250        2260        2270        2280
AGTACTAAAT GGAGAAAATT AGTAGATTTC AGAGAACTTA ATAAGAGAAC TCAAGACTTC

      2290        2300        2310        2320        2330        2340
TGGGAAGTTC AATTAGGAAT ACCACATGCC GCAGGGTTAA AAAAGAAAAA ATCAGTAACA
```

16 NOV.84- 29099

D F A

```
GTACTGGATG TGGGTGATGC ATATTTTTCA GTTCCCTTAG ATGAAGACTT CAGGAAGTAT

      2410       2420       2430       2440       2450       2460
ACTGCATTTA CCATACCTAG TATAAACAAT GAGACACCAG GGATTAGATA TCAGTACAAT

      2470       2480       2490       2500       2510       2520
GTGCTTCCAC AGGGATGGAA AGGATCACCA GCAATATTCC AAAGTAGCAT GACAAAAATC

      2530       2540       2550       2560       2570       2580
TTAGAGCCTT TTAGAAAACA AAATCCAGAC ATAGTTATCT ATCAATACAT GGATGATTTG

      2590       2600       2610       2620       2630       2640
TATGTAGGAT CTGACTTAGA AATAGGGCAG CATAGAACAA AAATAGAGGA GCTGAGACAA

      2650       2660       2670       2680       2690       2700
CATCTGTTGA GGTGGGGACT TACCACACCA GACAAAAAAC ATCAGAAAGA ACCTCCATTC

      2710       2720       2730       2740       2750       2760
CTTTGGATGG GTTATGAACT CCATCCTGAT AAATGGACAG TACAGCCTAT AGTGCTGCCA

      2770       2780       2790       2800       2810       2820
GAAAAAGACA GCTGGACTGT CAATGACATA CAGAAGTTAG TGGGAAAATT GAATTGGGCA

      2830       2840       2850       2860       2870       2880
AGTCAGATTT ACCCAGGGAT TAAAGTAAGG CAATTATGTA AACTCCTTAG AGGAACCAAA

      2890       2900       2910       2920       2930       2940
GCACTAACAG AAGTAATACC ACTAACAGAA GAAGCAGAGC TAGAACTGGC AGAAAACAGA

      2950       2960       2970       2980       2990       3000
GAGATTCTAA AAGAACCAGT ACATGGAGTG TATTATGACC CATCAAAAGA CTTAATAGCA

      3010       3020       3030       3040       3050       3060
GAAATACAGA AGCAGGGGCA AGGCCAATGG ACATATCAAA TTTATCAAGA GCCATTTAAA

      3070       3080       3090       3100       3110       3120
AATCTGAAAA CAGGAAAATA TGCAAGAACG AGGGGTGCCC ACACTAATGA TGTAAAACAA

      3130       3140       3150       3160       3170       3180
TTAACAGAGG CAGTGCAAAA AATAACCACA GAAAGCATAG TAATATGGGG AAAGACTCCT

      3190       3200       3210       3220       3230       3240
AAATTTAAAC TACCCATACA AAAGGAAACA TGGGAAACAT GGTGGACAGA GTATTGGCAA

      3250       3260       3270       3280       3290       3300
GCCACCTGGA TTCCTGAGTG GGAGTTTGTC AATACCCCTC CTTTAGTGAA ATTATGGTAC

      3310       3320       3330       3340       3350       3360
CAGTTAGAGA AAGAACCCAT AGTAGGAGCA GAAACGTTCT ATGTAGATGG GGCAGCTAGC

      3370       3380       3390       3400       3410       3420
AGGGAGACTA AATTAGGAAA AGCAGGATAT GTTACTAATA GAGGAAGACA AAAAGTTGTC

      3430       3440       3450       3460       3470       3480
ACCCTAACTG ACACAACAAA TCAGAAGACT GAGTTACAAG CAATTCATCT AGCTTTGCAG

      3490       3500       3510       3520       3530       3540
GATTCGGGAT TAGAAGTAAA TATAGTAACA GACTCACAAT ATGCATTAGG AATCATTCAA

      3550       3560       3570       3580       3590       3600
GCACAACCAG ATAAAAGTGA ATCAGAGTTA GTCAATCAAA TAATAGAGCA GTTAATAAAA

      3610       3620       3630       3640       3650       3660
```

221

Fig 22

```
      3670       3680       3690       3700       3710       3720
GTAGATAAAT TAGTCAGTGC TGGAATCAGG AAAGTACTAT TTTTAGATGG AATAGAT

      3730       3740       3750       3760       3770       3780
GCCCAAGATG AACATGAGAA ATATCACAGT AATTGGAGAG CAATGGCTAG TGATTTTAAC

      3790       3800       3810       3820       3830       3840
CTGCCACCTG TAGTAGCAAA AGAAATAGTA GCCAGCTGTG ATAAATGTCA GCTAAAAGGA

      3850       3860       3870       3880       3890       3900
GAAGCCATGC ATGGACAAGT AGACTGTAGT CCAGGAATAT GGCAACTAGA TTGTACACAT

      3910       3920       3930       3940       3950       3960
TTAGAAGGAA AAGTTATCCT GGTAGCAGTT CATGTAGCCA GTGGATATAT AGAAGCAGAA

      3970       3980       3990       4000       4010       4020
GTTATTCCAG CAGAAACAGG GCAGGAAACA GCATACTTTC TTTTAAAATT AGCAGGAAGA

      4030       4040       4050       4060       4070       4080
TGGCCAGTAA AAACAATACA TACAGACAAT GGCAGCAATT TCACCAGTAC TACGGTTAAG

      4090       4100       4110       4120       4130       4140
GCCGCCTGTT GGTGGGCGGG AATCAAGCAG GAATTTGGAA TTCCCTACAA TCCCCAAAGT

      4150       4160       4170       4180       4190       4200
CAAGGAGTAG TAGAATCTAT GAATAAAGAA TTAAAGAAAA TTATAGGCCA GGTAAGAGAT

      4210       4220       4230       4240       4250       4260
CAGGCTGAAC ATCTTAAGAC AGCAGTACAA ATGGCAGTAT TCATCCACAA TTTTAAAAGA

      4270       4280       4290       4300       4310       4320
AAAGGGGGGA TTGGGGGGTA CAGTGCAGGG GAAAGAATAG TAGACATAAT AGCAACAGAC

      4330       4340       4350       4360       4370       4380
ATACAAACTA AAGAATTACA AAAACAAATT ACAAAAATTC AAAATTTTCG GGTTTATTAC

      4390       4400       4410       4420       4430       4440
AGGGACAGCA GAGATCCACT TTGGAAAGGA CCAGCAAAGC TCCTCTGGAA AGGTGAAGGG

      4450       4460       4470       4480       4490       4500
GCAGTAGTAA TACAAGATAA TAGTGACATA AAAGTAGTGC CAAGAAGAAA AGCAAAGATC

      4510       4520       4530       4540       4550       4560
ATTAGGGATT ATGGAAAACA GATGGCAGGT GATGATTGTG TGGCAAGTAG ACAGGATGAG

      4570       4580       4590       4600       4610       4620
GATTAGAACA TGGAAAAGTT TAGTAAAACA CCATATGTAT GTTTCAGGGA AAGCTAGGGG

      4630       4640       4650       4660       4670       4680
ATGGTTTTAT AGACATCACT ATGAAAGCCC TCATCCAAGA ATAAGTTCAG AAGTACACAT

      4690       4700       4710       4720       4730       4740
CCCACTAGGG GATGCTAGAT TGGTAATAAC AACATATTGG GGTCTGCATA CAGGAGAAAG

      4750       4760       4770       4780       4790       4800
AGACTGGCAT CTGGGTCAGG GAGTCTCCAT AGAATGGAGG AAAAAGAGAT ATAGCACACA

      4810       4820       4830       4840       4850       4860
AGTAGACCCT GAACTAGCAG ACCAACTAAT TCATCTGTAT TACTTTGACT GTTTTTCAGA

      4870       4880       4890       4900       4910       4920
```

16 NOV 84- 29099

Fig 23

```
                                                                    CCTAGGTGTG AATATCAAGC

      4930        4940        4950        4960        4970        4980
AGGACATAAC AAGGTAGGAT CTGTACAATA CTTGGCACTA GCAGCATTAA TAACACCAAA

      4990        5000        5010        5020        5030        5040
AAAGATAAAG CCACCTTTGC CTAGTGTTAC GAAACTGACA GAGGATAGAT GGAACAAGCC

      5050        5060        5070        5080        5090        5100
CCAGAAGACC AAGGGCCACA GAGGGAGCCA CACAATGAAT GGACACTAGA GCTTTTAGAG

      5110        5120        5130        5140        5150        5160
GAGCTTAAGA ATGAAGCTGT TAGACATTTT CCTAGGATTT GGCTCCATGG CTTAGGGCAA

      5170        5180        5190        5200        5210        5220
CATATCTATG AAACTTATGG GGATACTTGG GCAGGAGTGG AAGCCATAAT AAGAATTCTG

      5230        5240        5250        5260        5270        5280
CAACAACTGC TGTTTATCCA TTTCAGAATT GGGTGTCGAC ATAGCAGAAT AGGCGTTACT

      5290        5300        5310        5320        5330        5340
CAACAGAGGA GAGCAAGAAA TGGAGCCAGT AGATCCTAGA CTAGAGCCCT GGAAGCATCC

      5350        5360        5370        5380        5390        5400
AGGAAGTCAG CCTAAAACTG CTTGTACCAC TTGCTATTGT AAAAAGTGTT GCTTTCATTG

      5410        5420        5430        5440        5450        5460
CCAAGTTTGT TTCACAACAA AAGCCTTAGG CATCTCCTAT GGCAGGAAGA AGCGGAGACA

      5470        5480        5490        5500        5510        5520
GCGACGAAGA CCTCCTCAAG GCAGTCAGAC TCATCAAGTT TCTCTATCAA AGCAGTAAGT

      5530        5540        5550        5560        5570        5580
AGTACATGTA ATGCAACCTA TACAAATAGC AATAGCAGCA TTAGTAGTAG CAATAATAAT

      5590        5600        5610        5620        5630        5640
AGCAATAGTT GTGTGGTCCA TAGTAATCAT AGAATATAGG AAAATATTAA GACAAAGAAA

      5650        5660        5670        5680        5690        5700
AATAGACAGG TTAATTGATA GACTAATAGA AAGAGCAGAA GACAGTGGCA ATGAGAGTGA

      5710        5720        5730        5740        5750        5760
AGGAGAAATA TCAGCACTTG TGGAGATGGG GGTGGAAATG GGGCACCATG CTCCTTGGGA

      5770        5780        5790        5800        5810        5820
TATTGATGAT CTGTAGTGCT ACAGAAAAAT TGTGGGTCAC AGTCTATTAT GGGGTACCTG

      5830        5840        5850        5860        5870        5880
TGTGGAAGGA AGCAACCACC ACTCTATTTT GTGCATCAGA TGCTAAAGCA TATGATACAG

      5890        5900        5910        5920        5930        5940
AGGTACATAA TGTTTGGGCC ACACATGCCT GTGTACCCAC AGACCCCAAC CCACAAGAAG

      5950        5960        5970        5980        5990        6000
TAGTATTGGT AAATGTGACA GAAAATTTTA ACATGTGGAA AAATGACATG GTAGAACAGA

      6010        6020        6030        6040        6050        6060
TGCATGAGGA TATAATCAGT TTATGGGATC AAAGCCTAAA GCCATGTGTA AAATTAACCC

      6070        6080        6090        6100        6110        6120
CACTCTGTGT TAGTTTAAAG TGCACTGATT TGCGGAATGC TACTAATACC AATAGTAGTA

      6130        6140        6150        6160        6170        6180
```

ATACCAATAG TAGTAGCGGG GAAATGATGA TGGAGAAAGG AGAGATAAAA AACTCCTCTT

D F A

6190 6200 6210 6220 6230 6240
TCAATATCAG CACAAGCATA AGAGGTAAGG TGCAGAAAGA ATATGCATTT TTTTATAAAC

6250 6260 6270 6280 6290 6300
TTGATATAAT ACCAATAGAT AATGATACTA CCAGCTATAC GTTGACAAGT TGTAACACCT

6310 6320 6330 6340 6350 6360
CAGTCATTAC ACAGGCCTGT CCAAAGGTAT CCTTTGAGCC AATTCCCATA CATTATTGTG

6370 6380 6390 6400 6410 6420
CCCCGGCTGG TTTTGCGATT CTAAAATGTA ATAATAAGAC GTTCAATGGA ACAGGACCAT

6430 6440 6450 6460 6470 6480
GTACAAATGT CAGCACAGTA CAATGTACAC ATGGAATTAG GCCAGTAGTA TCAACTCAAC

6490 6500 6510 6520 6530 6540
TGCTGTTGAA TGGCAGTCTA GCAGAAGAAG AGGTAGTAAT TAGATCTGCC AATTTCACAG

6550 6560 6570 6580 6590 6600
ACAATGCTAA AACCATAATA GTACAGCTGA ACCAATCTGT AGAAATTAAT TGTACAAGAC

6610 6620 6630 6640 6650 6660
CCAACAACAA TACAAGAAAA AGTATCCGTA TCCAGAGGGG ACCAGGGAGA GCATTTGTTA

6670 6680 6690 6700 6710 6720
CAATAGGAAA AATAGGAAAT ATGAGACAAG CACATTGTAA CATTAGTAGA GCAAAATGGA

6730 6740 6750 6760 6770 6780
ATGCCACTTT AAAACAGATA GCTAGCAAAT TAAGAGAACA ATTTGGAAAT AATAAAACAA

6790 6800 6810 6820 6830 6840
TAATCTTTAA GCAATCCTCA GGAGGGGACC CAGAAATTGT AACGCACAGT TTTAATTGTG

6850 6860 6870 6880 6890 6900
GAGGGGAATT TTTCTACTGT AATTCAACAC AACTGTTTAA TAGTACTTGG TTTAATAGTA

6910 6920 6930 6940 6950 6960
CTTGGAGTAC TGAAGGGTCA AATAACACTG AAGGAAGTGA CACAATCACA CTCCCATGCA

6970 6980 6990 7000 7010 7020
GAATAAAACA ATTTATAAAC ATGTGGCAGG AAGTAGGAAA AGCAATGTAT GCCCCTCCCA

7030 7040 7050 7060 7070 7080
TCAGCGGACA AATTAGATGT TCATCAAATA TTACAGGGCT GCTATTAACA AGAGATGGTG

7090 7100 7110 7120 7130 7140
GTAATAACAA CAATGGGTCC GAGATCTTCA GACCTGGAGG AGGAGATATG AGGGACAATT

7150 7160 7170 7180 7190 7200
GGAGAAGTGA ATTATATAAA TATAAAGTAG TAAAAATTGA ACCATTAGGA GTAGCACCCA

7210 7220 7230 7240 7250 7260
CCAAGGCAAA GAGAAGAGTG GTGCAGAGAG AAAAAAGAGC AGTGGGAATA GGAGCTTTGT

7270 7280 7290 7300 7310 7320
TCCTTGGGTT CTTGGGAGCA GCAGGAAGCA CTATGGGCGC ACGGTCAATG ACGCTGACGG

7330 7340 7350 7360 7370 7380
TACAGGCCAG ACAATTATTG TCTGGTATAG TGCAGCAGCA GAACAATTTG CTGAGGGCTA

7390 7400 7410 7420 7430 7440

AAGGCGCA ACAUCATCTG TTGCAACTCA CAGTCTGGGG CATCAAGCAG CTCCAGGCAA
D F A

```
      7450        7460        7470        7480        7490        7500
GAATCCTGGC TGTGGAAAGA TACCTAAAGG ATCAACAGCT CCTGGGGATT TGGGGTTGCT

      7510        7520        7530        7540        7550        7560
CTGGAAAACT CATTTGCACC ACTGCTGTGC CTTGGAATGC TAGTTGGAGT AATAAATCTC

      7570        7580        7590        7600        7610        7620
TGGAACAGAT TTGGAATAAC ATGACCTGGA TGGAGTGGGA CAGAGAAATT AACAATTACA

      7630        7640        7650        7660        7670        7680
CAAGCTTAAT ACATTCCTTA ATTGAAGAAT CGCAAAACCA GCAAGAAAAG AATGAACAAG

      7690        7700        7710        7720        7730        7740
AATTATTGGA ATTAGATAAA TGGGCAAGTT TGTGGAATTG GTTTAACATA ACAAATTGGC

      7750        7760        7770        7780        7790        7800
TGTGGTATAT AAAAATATTC ATAATGATAG TAGGAGGCTT GGTAGGTTTA AGAATAGTTT

      7810        7820        7830        7840        7850        7860
TTGCTGTACT TTCTATAGTG AATAGAGTTA GGCAGGGATA TTCACCATTA TCGTTTCAGA

      7870        7880        7890        7900        7910        7920
CCCACCTCCC AACCCCGAGG GGACCCGACA GGCCCGAAGG AATAGAAGAA GAAGGTGGAG

      7930        7940        7950        7960        7970        7980
AGAGAGACAG AGACAGATCC ATTCGATTAG TGAACGGATC CTTAGCACTT ATCTGGGACG

      7990        8000        8010        8020        8030        8040
ATCTGCGGAG CCTTGTGCCT CTTCAGCTAC CACCGCTTGA GAGACTTACT CTTGATTGTA

      8050        8060        8070        8080        8090        8100
ACGAGGATTG TGGAACTTCT GGGACGCAGG GGGTGGGAAG CCCTCAAATA TTGGTGGAAT

      8110        8120        8130        8140        8150        8160
CTCCTACAGT ATTGGAGTCA GGAACTAAAG AATAGTGCTG TTAGCTTGCT CAATGCCACA

      8170        8180        8190        8200        8210        8220
GCCATAGCAG TAGCTGAGGG GACAGATAGG GTTATAGAAG TAGTACAAGG AGCTTGTAGA

      8230        8240        8250        8260        8270        8280
GCTATTCGCC ACATACCTAG AAGAATAAGA CAGGGCTTGG AAAGGATTTT GCTATAAGAT

      8290        8300        8310        8320        8330        8340
GGGTGGCAAG TGGTCAAAAA GTAGTGTGGT TGGATGGCCT ACTGTAAGGG AAAGAATGAG

      8350        8360        8370        8380        8390        8400
ACGAGCTGAG CCAGCAGCAG ATGGGGTGGG AGCAGCATCT CGAGACCTGG AAAAACATGG

      8410        8420        8430        8440        8450        8460
AGCAATCACA AGTAGCAATA CAGCAGCTAC CAATGCTGCT TGTGCCTGGC TAGAAGCACA

      8470        8480        8490        8500        8510        8520
AGAGGAGGAG GAGGTGGGTT TTCCAGTCAC ACCTCAGGTA CCTTTAAGAC CAATGACTTA

      8530        8540        8550        8560        8570        8580
CAAGGCAGCT GTAGATCTTA GCCACTTTTT AAAAGAAAAG GGGGGACTGG AAGGGCTAAT

      8590        8600        8610        8620        8630        8640
TCACTCCCAA CGAAGACAAG ATATCCTTGA TCTGTGGATC TACCACACAC AAGGCTACTT

      8650        8660        8670        8680        8690        8700
```

DNA

```
CCCTGATTGG CAGAACTACA CACCAGGGCC AGGGGTCAGA TATCCACTGA CCTTTGGATC

      8710        8720        8730        8740        8750        8760
GTGCTACAAG CTAGTACCAG TTGAGCCAGA TAAGGTAGAA GAGGCCAATA AAGGAGAGAA

      8770        8780        8790        8800        8810        8820
CACCAGCTTG TTACACCCTG TGAGCCTGCA TGGAATGGAT GACCCTGAGA GAGAAGTGTT

      8830        8840        8850        8860        8870        8880
AGAGTGGAGG TTTGACAGCC GCCTAGCATT TCATCACGTG CCCCGAGAGC TGCATCCGGA

      8890        8900        8910        8920        8930        8940
GTACTTCAAG AACTGCTGAC ATCGAGCTTG CTACAAGGGA CTTTCCGCTG GGGACTTTCC

      8950        8960        8970        8980        8990        9000
AGGGAGGCGT GGCCTGGGCG GAACTGGGGA GTGGCGAGCC CTCAGATGCT GCATATAAGC

      9010        9020        9030        9040        9050        9060
AGCTGCTTTT TGCCTGTACT GGGTCTCTCT GGTTAGACCA GATTTGAGCC TGGGAGCTCT

      9070        9080        9090        9100           0           0
CTGGCTAACT AGGGAACCCA CTGCTTAAGC CTCAATAAAG CTT
```

Fig 26